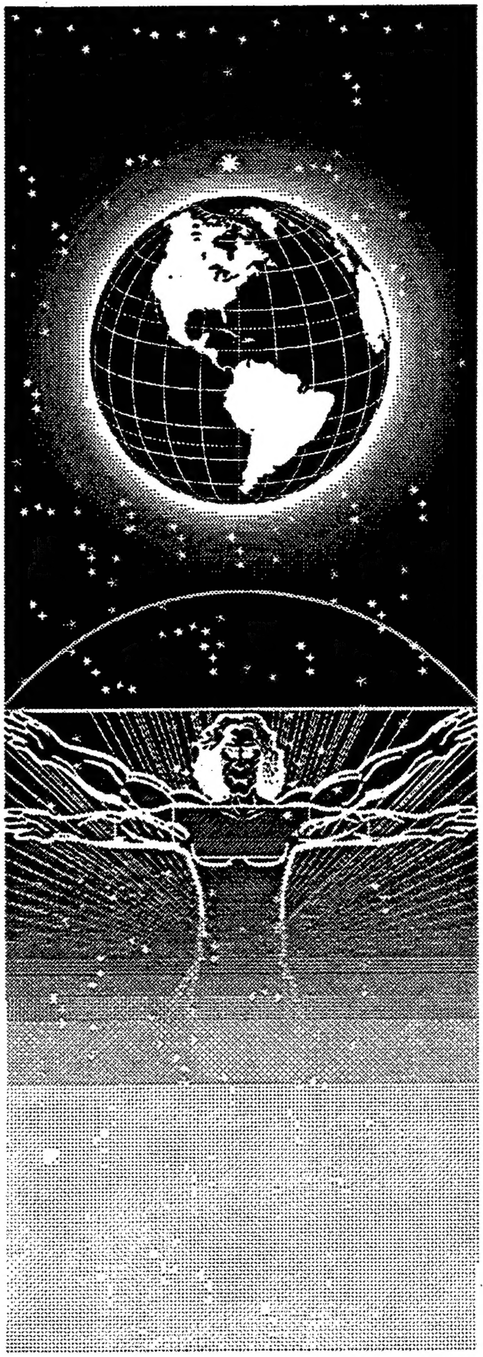
AFRL-HE-WP-TR-2001-0116



UNITED STATES AIR FORCE RESEARCH LABORATORY

AIDING THE INTELLIGENCE ANALYST: FROM PROBLEM DEFINITION TO DESIGN CONCEPT EXPLORATION

Emily S. Patterson
David D. Woods
David Tinapple

COGNITIVE SYSTEMS ENGINEERING LABORATORY
INSTITUTE FOR ERGONOMICS
THE OHIO STATE UNIVERSITY
COLUMBUS OH 43210

Emilie M. Roth

ROTH COGNITIVE ENGINEERING
89 RAWSON ROAD
BROOKLINE MA 02445-4509

John M. Finley

NATIONAL AIR INTELLIGENCE CENTER
WRIGHT-PATTERSON AFB OH 45433-7106

Gilbert G. Kuperman

HUMAN EFFECTIVENESS DIRECTORATE
CREW SYSTEM INTERFACE DIVISION
WRIGHT-PATTERSON AFB OH 45433-7022

MARCH 2001

INTERIM REPORT FOR THE PERIOD 1 FEBRUARY 2000 TO 30 MARCH 2001

Human Effectiveness Directorate
Crew System Interface Division
2255 H Street
Wright-Patterson AFB OH 45433-7022

Approved for public release; distribution is unlimited.

NOTICES

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-2001-0116

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

MARIS M. VIKMANIS
Chief, Crew System Interface Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** March 2001	**3. REPORT TYPE AND DATES COVERED** Interim Report -- 1 February 2000 to 30 March 2001

4. TITLE AND SUBTITLE Aiding the Intelligence Analyst: From Problem Definition to Design Concept Exploration	**5. FUNDING NUMBERS** C: F33615-98-D-6000 PE: 62202F PR: 7184 TA: 10 WU: 01
6. AUTHOR(S) Patterson, E.S.*, Woods, D.D.*, Tinapple, D. *, Roth, E.M. **, Finley, J. M.***, Kuperman, G.G. ****	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) * The Ohio State University, Columbus OH 34210 ** Roth Cognitive Engineering, 89 Rawson Road, Brookline MA 02445-4509 *** National Air Intelligence Center, Wright-Patterson AFB, OH 45433-7106 **** Air Force Research Laboratory, Wright-Patterson AFB, OH 45433-7022	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory Human Effectiveness Directorate Crew System Interface Division Air Force Materiel Command Wright-Patterson AFB, OH 45433-7022	**10. SPONSORING/MONITORING AGENCY REPORT NUMBER** AFRL-HE-WP-TR-2001-0116

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.	**12b. DISTRIBUTION CODE**

13. ABSTRACT *(Maximum 200 words)*

This report details a complete, beginning-to-end Cognitive Systems Engineering (CSE) project tackling the challenges of conducting intelligence analysis under the condition of data overload. We first reviewed and synthesized the research base on data overload from multiple complex, high-consequence settings like nuclear power generation. We conducted a study to identify the aspects of this research base that applied to intelligence analysis as well as unique challenges. We observed expert intelligence analysts conducting an analysis and we identified challenging tasks that leave analysts vulnerable to making inaccurate statements. We developed modular design concepts, or "design seeds," that could be incorporated into both ongoing and future design efforts.

14. SUBJECT TERMS cognitive engineering, cognitive task analysis, data overload, design seed, inferential analysis, information retrieval, information visualization, intelligence analysis, simulation study	**15. NUMBER OF PAGES** 107
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNL

This page left blank intentionally.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

This page left blank intentionally.

PART I. INTRODUCTION

This report details a complete, beginning-to-end Cognitive Systems Engineering (CSE) project tackling the challenges of conducting intelligence analysis under the condition of data overload. We first reviewed and synthesized the research base on data overload from multiple complex, high-consequence settings like nuclear power generation. Then, leveraging this research base and previous experience in conducting Cognitive Task Analysis (CTA), we conducted a study to identify the aspects of this research base that applied to intelligence analysis as well as unique challenges. We observed expert intelligence analysts conducting an analysis on a selected unclassified scenario, the 1996 Ariane 501 rocket launch failure, with a baseline set of tools that supported keyword search, browsing, and word processing in an investigator-constructed database. From this study, we identified challenging tasks in intelligence analysis that leave analysts vulnerable to making inaccurate statements in briefings when they are working in a new topic area and are under short deadline constraints. In parallel, we identified limitations of the baseline tools in addressing these vulnerabilities that pointed to ideas for new design directions. In addition, the study findings were translated into objective criteria for evaluating the usefulness of any effort aimed at reducing data overload.

In the final phase of the project, we shifted from an emphasis on problem definition to an emphasis on developing modular design concepts, or "design seeds," that could be incorporated into both ongoing and future design efforts. The design seeds were instantiated as animated fly-through mockups, or "Animocks," so that feasibility of certain usability considerations, such as the ability to display data in parallel in an easily interpretable form, could be explored without being forced into committing to a particular design. With Animocks, commitment to a particular hardware infrastructure, visualization instantiation, or combination of design seeds is lessened and there is greater flexibility to incorporate feedback about the usefulness of a design concept in addressing data overload. In addition, the design seeds are conceptually modular and based in challenging scenarios, which enables generalization of concepts across design projects and domains. With this strategy, we can better address one of the primary challenges of a research and development program, which is to develop research bases that can be translated into fieldable systems in multiple settings to prevent continuously engaging in individual, one-off design endeavors without learning how to improve systems over time. We believe that complementarity of research and design efforts and accompanying cross-stimulation is the main characteristic of effective Cognitive Systems Engineering Research and Development (R&D) programs.

Many are interested in R&D at the intersection of people, technology and work. R&D at the intersection of people, technology and work is currently a world divided and hobbled. Research results seem irrelevant or difficult to translate to a specific design project. Design is a cumbersome process of trial and error, with little cross-fertilization and learning across design projects. In the rush to the deadline, we often create expensive systems that are not viewed as useful by practitioners, and often never even used in a real setting.

The standard metaphor and organizational construct is a pipeline from basic research to more applied research to design. This pipeline metaphor has failed to create effective interconnections and cross-stimulation between research and design activities. True design innovation is difficult to achieve within funding structures based on this metaphor.

In this report, we provide and illustrate with a complete, beginning-to-end example an alternative model for R&D that we used to stimulate true innovation and create positive synergies between scientific advancement and practice-centered design. The strategy of complementarity between research and design upon which it is based is foundational to the intent behind the label Cognitive Systems Engineering (and related labels like distributed cognition and naturalistic decision making) as an alternative to traditional disciplinary approaches. Rather than the pipeline metaphor, we use a metaphor of interlocking gears to describe synchronization of multiple parallel cycles of learning and development that operate at different time scales. Interlocking these cycles is a difficult challenge – a challenge in producing organizational frameworks and supporting mechanisms to create and extend innovation.

Before describing the case study in detail, we now describe our general philosophy behind our approach and highlight how it is different than traditional views in that:

1) We view science and design as complementary, mutually reinforcing activities,
2) We view design as an iterative "bootstrap" process rather than a linear process,
3) We use prototypes as tools for discovery to probe the interaction of people, technology and work rather than as partially refined final designs,
4) We separate out learning on three levels throughout the design process: understanding the challenges in a domain, determining what would be useful aids to domain practitioners, and improving the usability of artifacts, and
5) We focus our R&D investments on the "usefulness" level of design in order to target leverage points that will have the most impact on the end practitioners' ability to meet domain challenges.

1.1 Complementarity of Science and Design

"It is...the fundamental principle of cognition that the universal can be perceived only in the particular, while the particular can be thought of only in reference to the universal."

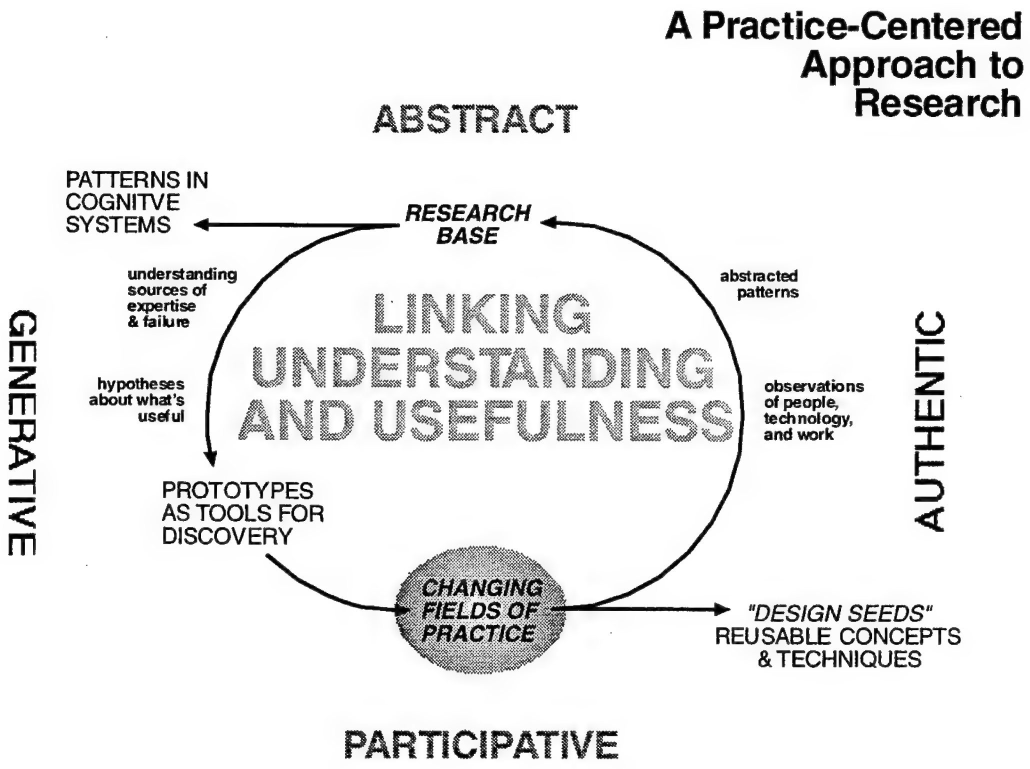
-- Cassirer 1923/1953, p. 86

"If we truly understand cognitive systems, then we must be able to develop designs that enhance the performance of operational systems; if we are to enhance the performance of operational systems, we need conceptual looking glasses that enable us to see past the unending variety of technology and particular domains."

-- Woods and Sarter, 1993

There is a common misconception that research and design are separate, independent activities. This division is artificial. It is possible that the skills required to perform research and design are distinct, but the activities themselves are mutually reinforcing. As depicted in Figure 1, every design that is introduced into a field of practice embodies a hypothesis about what would be a useful artifact. The introduction of this design serves as a natural experiment where observations of the impact of the design on a field of practice can be abstracted to patterns that gain authenticity when viewed in cognitive systems in multiple settings. Similarly, every scientific experiment requires design of the artifacts used to support the practitioners.[1] Field settings are natural laboratories for longer term learning, and, at the same time, they are fields of practice where technological and organizational interventions are introduced in attempts to improve performance.

[1] Note that in most experimental psychology laboratory studies, this means that there are no artifacts to support cognition. A common misconception is that true science requires the study of a phenomenon in isolation of its surrounding environment (e.g., studies of human memory that do not allow study participants access to paper and pencil). The argument that science must be conducted in a "no world" environment in order to facilitate generalization is not convincing. In field research, the complexities of the field must be maintained in the unit of study because this is where the phenomena of interest occur. The challenge is to avoid becoming lost in the details of each unique setting– to characterize regularities in scientific terms that generalize across multiple settings.


A Practice-Centered Approach to Research
ABSTRACT
PATTERNS IN COGNITVE SYSTEMS
RESEARCH BASE
understanding sources of expertise & failure
abstracted patterns
GENERATIVE
LINKING UNDERSTANDING AND USEFULNESS
hypotheses about what's useful
observations of people, technology, and work
AUTHENTIC
PROTOTYPES AS TOOLS FOR DISCOVERY
CHANGING FIELDS OF PRACTICE
"DESIGN SEEDS" REUSABLE CONCEPTS & TECHNIQUES
PARTICIPATIVE

© 2000 Christoffersen, Woods, and Malin

Figure 1. Complementarity Defines Cognitive Systems Engineering

Two coordinated strands define complementarity. In one strand (Figure 2), inquiry is directed at capturing phenomena, abstracting patterns and discovering the forces that produce those phenomena despite the surface variability of different technology and different settings. In this sense, effective research develops a book of "patterns" as a generic but relevant research base.

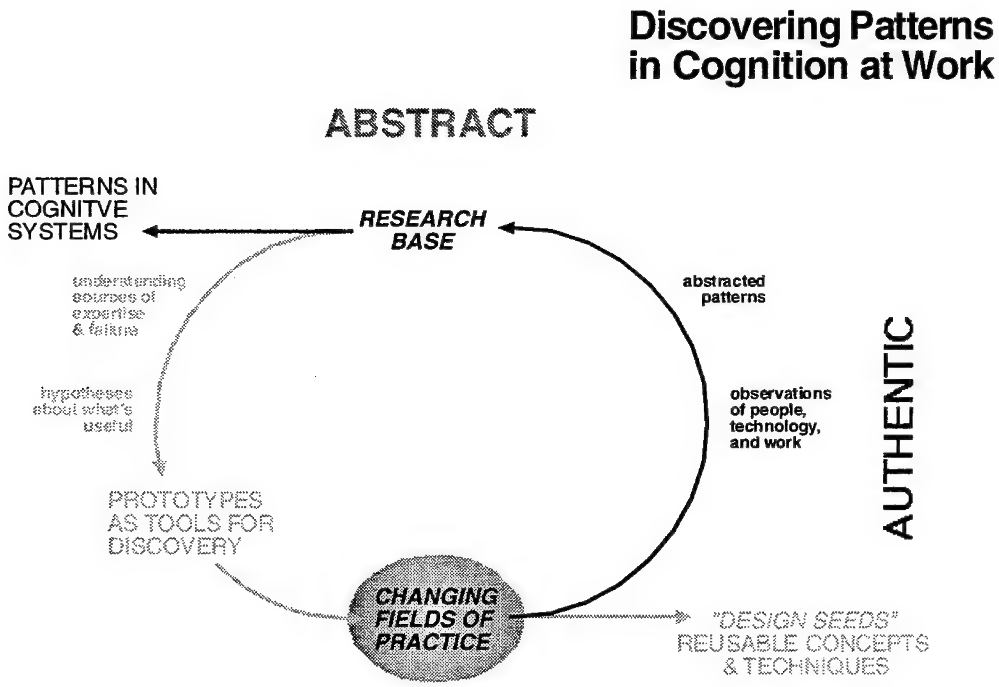
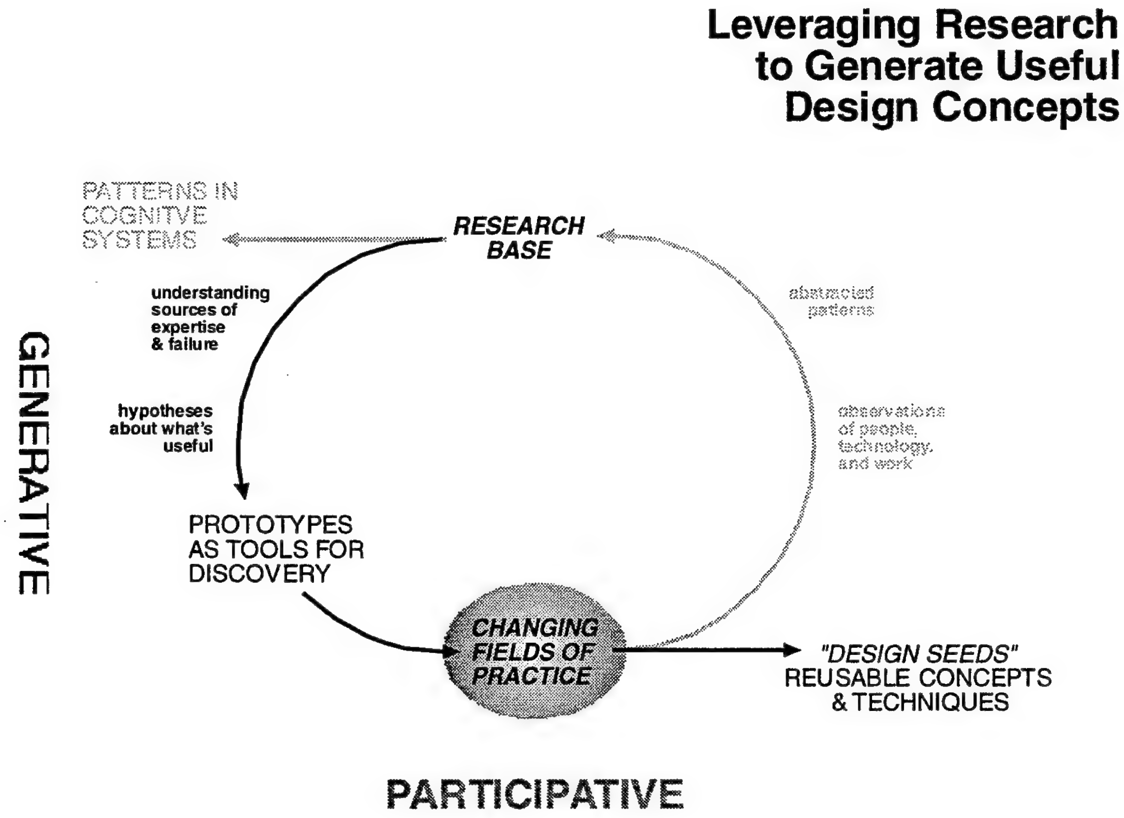


Figure 2. Discovering Patterns in Cognition at Work.
Observing, abstracting, explaining phenomena at the intersection of people, technology and work.

But the challenge of stimulating innovation goes further. A second strand of processes is needed that link this tentative understanding to the process of discovering what would be useful (Figure 3). Success occurs when "reusable" (that is, tangible but relevant to multiple settings) design concepts and techniques are created to "seed" the systems development cycle.


Leveraging Research to Generate Useful Design Concepts
PATTERNS IN COGNITVE SYSTEMS
RESEARCH BASE
understanding sources of expertise & failure
hypotheses about what's useful
PROTOTYPES AS TOOLS FOR DISCOVERY
CHANGING FIELDS OF PRACTICE
"DESIGN SEEDS" REUSABLE CONCEPTS & TECHNIQUES
abstracted patterns
observations of people, technology, and work
GENERATIVE
PARTICIPATIVE

© 2000 Christoffersen, Woods, and Malin

Figure 3. Leveraging Research to Generate Useful Design Concepts. Generating reusable concepts about what would be useful to seed development.

In the end, innovation is stimulated both through creation of possible futures and reflection about the effects of those while the commitment to any particular object is relaxed and the limited horizon of development cycles is stretched. The combination creates a complementary cycle of learning and development. Advancing our understanding abstracts patterns and phenomena from observations of the interplay of people, technology and work and develops explanations for the appearance of these patterns across different fields of practice. This cycle seeks to discover performance-related issues within each given setting and to develop hypotheses about what may be useful in response to these issues. Aiding concepts are embodied in prototypes as part of a continuing learning and discovery process. Over time, the result is a generically defined set of concepts and techniques that can seed development in multiple specialized areas where the relevant performance issues play out.

An effective balance generates two types of advances, each as tentative syntheses of what we think we know about the interplay of people, technology and work.

The research base is seen as patterns abstracted across different unique settings, patterns that are in need of explanation and concepts that could explain these observations. As Hutchins (1992) put it, "There are powerful regularities to be described at a level of analysis that transcends the details of the specific domain. It is not possible to discover these regularities without understanding the details of the domain, but the regularities are not about the domain specific details, they are about the nature of human cognition in human activity."

The second product of an effective balance would be the ability to capture and share design "seeds" – concepts and techniques about what would be useful to advance cognition and collaboration at work. These are seeds in the sense that they stimulate innovation in different specific settings. "If we are to enhance the performance of operational systems, we need conceptual looking glasses that enable us to see past the unending variety of technology and particular domains" (Woods and Sarter, 1993). To achieve this complementarity, usefulness, i.e., criteria that new systems enhance performance in context becomes a criterion for research (can it effectively seed and leverage development in more than a specific case).

In coordinating these processes four kinds of activities go on. Fields of practice are the primary focus. Authentic samples of what it means to practice in that field of activity and how the organizational dynamics pressure or support practice stimulate the process of learning. However, the observer will quickly become lost in the detail of particular settings at particular points in time with particular technological objects unless they can compare and contrast settings over time to abstract patterns and produce candidate explanations for the basic patterns. The third component is generative – in studying the interaction of people, technology and work across fields of practice we must generate or discover new ideas, including explanations for the phenomena and patterns observed, but more critically, new hypotheses about what would be useful to probe the field of practice, test our tentative understanding, and to seed upcoming development cycles. In the final analysis the activity is participative as we work with practitioners in these field of activities to understand how they adapt to the pressures and demands of the field of activity.

The two half cycles (Figures 2 and 3) are inter-dependent, not separate. The point of the processes of observation, abstraction and explanation is to find the essential factors under the surface variability. In other words, the test of understanding is the ability to anticipate the impacts of technological change. The ultimate risk for researchers is to acknowledge that they are part of the process under study:

- to participate in the struggle of envisioning with other stakeholders and
- to acknowledge their role as designers -- the development of tools that make us smart or dumb.

The ultimate test for the designer is:

- to risk abstraction and acknowledge their prototypes as hypotheses at empirical jeopardy.

In a practice-centered process, we face challenges related to four basic values:

- Transcending **limits to authenticity** to capture how the strategies and behavior of people are adapted to the constraints and demands of fields of practice,
- Meeting the **challenge of abstraction** to find patterns behind the surface variability,
- **Sparking inventiveness** to discover new ways to use technological possibilities to enhance human performance, to identify leverage points, and to minimize unanticipated side effects, and
- **Creating future possibilities** as participants with other stakeholders and problem holders in that field of practice.

With this view of complementarity of research and design, progress in one of the areas feeds into the other. For example, if a system design succeeds in reducing error and supporting or enhancing expertise, we can expand or refine our models of error and expertise for practitioners under similar demands across multiple settings. Similarly, if we develop predictive models of how interventions will impact error and expertise in a domain, these models can form underlying concepts for system designs. All designs embody concepts of what will be useful to practitioners, whether only implicitly or explicitly. It is through the careful examination of how deployed systems impact error and expertise that we can improve our ability to predict and embody the predictions in deployed systems.

1.2 Understanding, Usefulness, and Usability

Discovery of what would be useful occurs in the research cycle because development also functions as an opportunity to learn. Artifacts are not just objects; they are hypotheses about the interplay of people, technology and work. With this perspective of complementarity between research and design, prototypes function as tools for discovery to probe the interaction of people, technology and work and to test the hypothesized, envisioned impact of technological change.

Many commentators note that design begins with analysis and problem definition. This phase is concerned with understanding the nature of the problem to be solved and the field of practice, the domain, in which it resides. Design activity then progresses to a divergent phase of ideation, ending in the selection of a single idea or set of ideas. Ideas, at this point, are means of achieving the desired ends; they are how the product will work. The final product is then produced through a gradual refinement phase, during which it is evaluated to ensure it meets its requirements. Each of these three phases builds on the others: analysis supports ideation, and the final product is an embodiment of the selected idea and is evaluated based on how it solves the problem identified in the analysis phase.

Rather than view design as this phased, sequential process, we believe that the design process draws on research bases and past experience in a creative, iterative fashion. With this view, the traditional stages in design are transformed to parallel levels of knowledge bases from which we can draw on opportunistically during the iterative design process. Rather than phases, the traditional stages represent tracks that proceed in parallel but produce different "products": a model of error and expertise in the domain, aiding concepts for what will be useful to practitioners, and the fielded system (Figure 4).

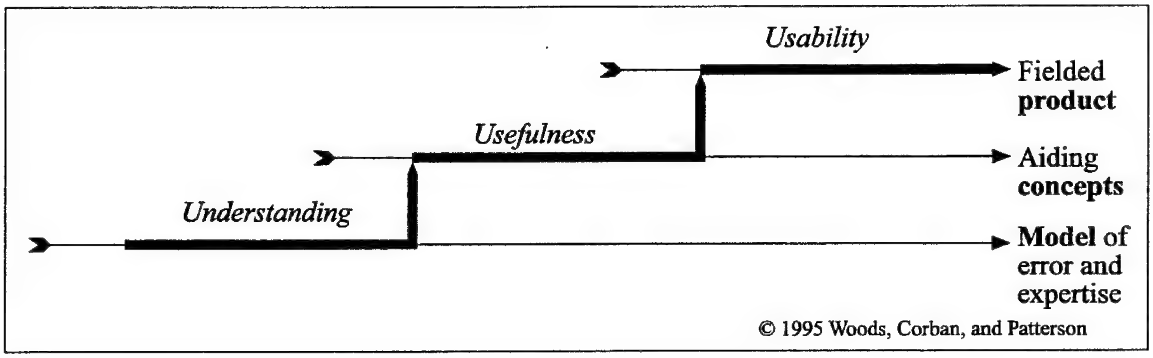


Figure 4. Three Parallel Design Tracks

1.2.1 Understanding

For practice-centered design, the problem definition is more than merely learning about the field of practice and talking to the practitioners; the designer must understand the nature of errors that occur and how experienced practitioners develop and maintain expertise. By understanding the demands that practitioners must meet in order to be successful, we can identify constraints on productive design directions. For example, if we know that intelligence analysts should never miss evidence that a nuclear weapon has been tested, system designs that hide information that is infrequently reviewed are not likely to meet the demands.

1.2.2 Usefulness

The usefulness track is where innovation occurs; during this phase, designers generate creative ideas for what might help users. It is intertwined with domain modeling, because testing ideas adds new knowledge to the designer's model of error and expertise, which can in turn lead to new ideas for how to aid performance. At this level, evaluations are aimed at the underlying generalizable aiding concept rather than the specific implementation. In order to do this, evaluation scenarios need to be crafted in ways to discover further requirements about what might be useful rather than refining the product to be more usable and consistent with existing infrastructures.

1.2.3 Usability

During prototype refinement, a series of decisions are made that continually narrow an idea into a fieldable product. At the same time, there is a growing resource and psychological commitment to a single concept. Design activity concerns making commitments to how specific aspects of the product will look or work. The depth of activity is often impasse driven, i.e., additional information search, evaluation, or consulting with human-computer interaction (HCI) specialists occurs when the design team confronts some impasse or gap in their knowledge. HCI specialists can draw on knowledge of principles and techniques to enhance usability. Usability tests are designed around cases that instantiate target scenarios that were derived from the understanding of the demands of the field of practice.

In order to advance as a science, we also can view each design process as an opportunity to add back to the three research bases. Any embodiment of a design concept, whether a crude prototype, refined prototype, or released system, affords the opportunity to add back to these research bases simultaneously: by improving our models of error and expertise, by discovering requirements for what would be useful, and by identifying principles and techniques to enhance usability.

1.3 Balancing Investments in Understanding, Usefulness, and Usability

Figure 5 integrates the three parallel levels of understanding, usefulness, and usability and emphasizes the complementarity between research and design. On the design side of the diamond, the understanding of the challenges in a work setting determines constraints on areas of the design space that might prove useful. Design concepts embody ideas of what will be useful to a practitioner. In order to have the design concept work effectively, it also needs to be easy to use and learn -- the usability level.

We believe that research efforts should be complementary with design efforts in order to ground research in important problems and make the research useful to real-world practitioners. Therefore, at the understanding level, research can be used to better understand the challenges of a work setting, which is directly valuable to determining constraints on design directions to pursue. At the usefulness level, scenarios in research studies are constructed to probe the domain challenges. Scenarios are designed to reveal how well an artifact supports a practitioner in meeting a domain challenge. In order to use these scenarios in a research study, they need to be converted to cases that need to be "usable" in the sense of understandable by study participants and targeted to allow investigators to unambiguously interpret the behavior of study participants in relation to a specific target for research.

The diamond shape illustrates our belief that an emphasis on the usefulness level is important in order to have a balanced R&D "investment portfolio." We believe that generating and evaluating generalizable aiding concepts through carefully crafted scenarios will allow us to make the most effective progress. Although it is important to invest in an understanding of the demands of a domain in order to reduce the risk of designing systems that are not useful, and although it is important to refine the usability of a product before deploying it in order to prevent a good concept from being rejected "in the field" for superficial reasons, we believe that the greatest investment should be in ideation of multiple concepts with an eye to throwing away ideas that are not useful. This portfolio is designed to minimize the risks and unnecessary expense of repeatedly building systems that do not support users in the field in meeting the most important domain challenges, such as analysis under data overload.

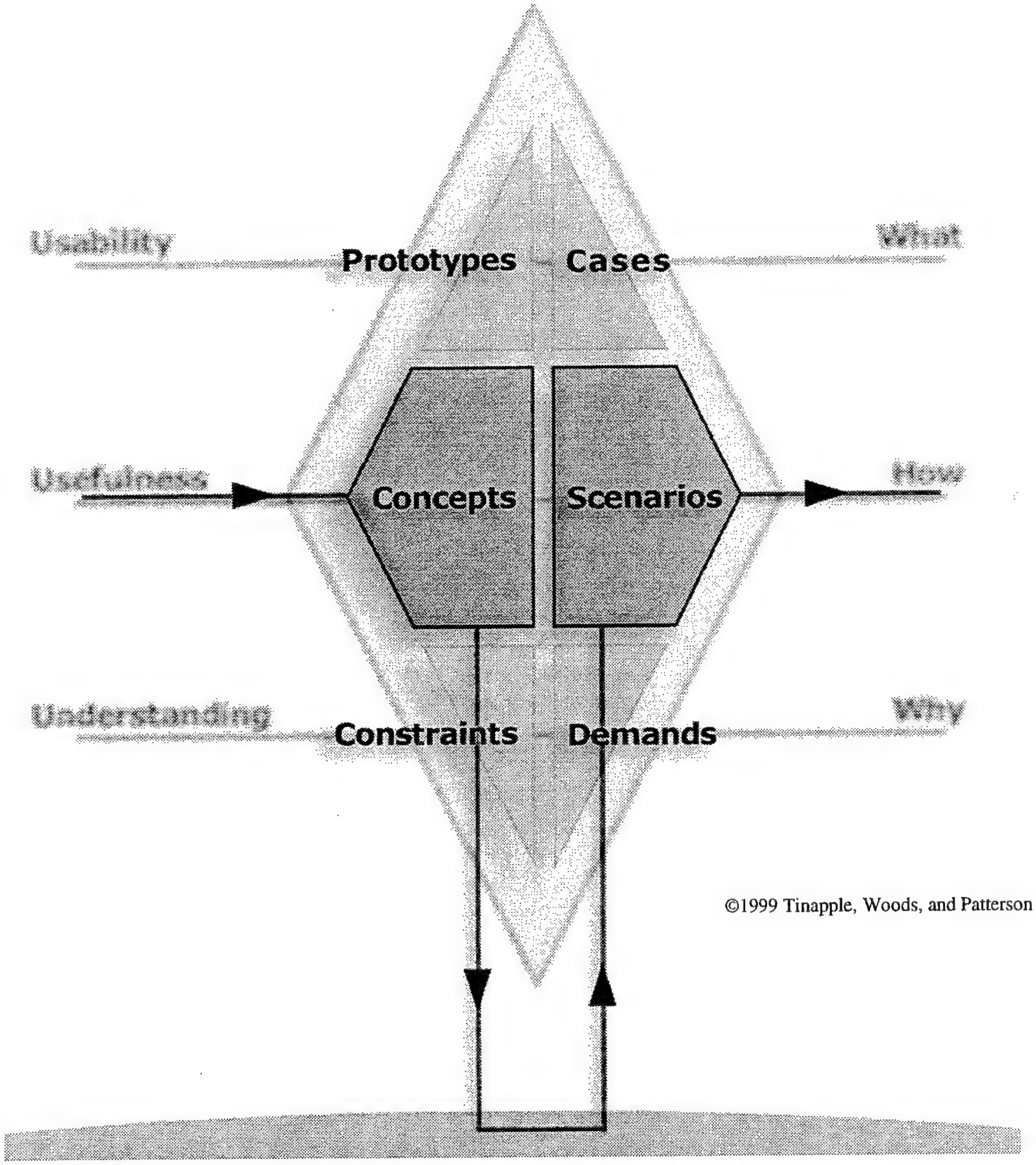


Figure 5. Research and Design at Three Levels of Abstraction

PART II. DIAGNOSING THE DATA OVERLOAD PROBLEM

We now detail an example of a complete, beginning-to-end Cognitive Systems Engineering (CSE) project tackling the challenges of conducting intelligence analysis under the condition of data overload where we treated research and design as complementary, cross-fertilizing activities. In this project, we

- Diagnosed the data overload problem and synthesized existing approaches to combat data overload,
- Calibrated the aspects of the research base on data overload that are relevant to intelligence analysis and identified aspects unique to intelligence analysis through observations of expert analysts on a simulated Quick Reaction Task (QRT),
- Identified challenging tasks that leave intelligence analysts vulnerable to making inaccurate statements when they are working in a new topic area and are under short deadline constraints,
- Identified limitations of baseline tools addressing vulnerabilities that point the way to new directions for design approaches to combat data overload
- Generated objective criteria for evaluating the usefulness of design concepts addressing data overload, and
- Developed modular design concepts in animated fly-through mock-ups, Animocks, that can be explored in future studies of what would be useful to inferential analysis under data overload as well as directly incorporated into ongoing and future design of fielded systems.

2.1 The Presentation of the Data Overload Challenge

The challenge presented to the team was to develop information visualizations that would help intelligence analysts deal with the avalanche of electronic, textual data that is available for them to use in generating a coherent response to a question. Specifically, there was interest in helping analysts find appropriate information to formulate answers to questions outside their immediate expertise under a tight deadline such as 24 hours, often referred to in the intelligence community as Quick Reaction Tasks (QRTs).

In one sense, the problem seems paradoxical because all intelligence analysts agree that access to more data ought to be a benefit. However, the benefit in principle has not been matched by the benefit in practice. The sheer volume of the data creates a situation where it is difficult to determine where to look in the data field, it becomes easy to miss critical information, and determining the significance of data in relation to the ongoing context is challenging.

Although data overload, or perhaps more descriptively data avalanche, in intelligence analysis is a compelling problem for real practitioners who really have to do important work and need immediate help, we felt that, in order to make more than incremental progress, it was important to step back and better characterize the problem before committing to a particular design direction. Therefore, rather than immediately constructing and evaluating information visualizations for text summarizer algorithms as originally requested, we delved into Cognitive Systems Engineering (CSE) research bases and our team members' past experiences with researching and designing parameter data displays in space shuttle mission control (Malin et al., 1991; Thronesbery, Christoffersen, Malin, 1999), display and mode proliferation on aviation flight decks (Sarter and Woods, 1992; Billings, 1996), agent explanations and alarms in anesthesiology (Cook and Woods, 1996; Johannesen, Cook, and Woods, 1994), distributed military command and control (Shattuck and Woods, 1997), navigation in computer displays and spreadsheets (Woods and Watts, 1997), and alarm overload in nuclear power generation (Woods, 1995a) with the following questions:

- What is the definition of data overload?
- Why is data overload so difficult to address?
- Why have new waves of technology exacerbated, rather than resolved, data overload?
- How are people able to cope with data overload?

2.2 Defining Data Overload

An important early step was to translate the cognitive tasks, intelligence analysis domain, and data overload problem into scientific terms that would allow us to leverage relevant Cognitive Systems Engineering research bases. Therefore, we were able to determine relatively quickly that:

- The main cognitive task in intelligence analysis is inferential analysis, which involves determining the best explanation for uncertain, often contradictory and incomplete data. The inferential analysis task could be defined as abductive reasoning (Josephson and Josephson, 1994) in the sense that the analytic product is always contestable because of the uncertainty, but that certain conclusions and analytic processes could be argued to be better than others and recognized as such by experts.

- Another cognitive framing of intelligence analysis could be that of a supervisory controller (the intelligence analyst) monitoring a process (national technological and human processes/capabilities). The main difference between traditional supervisory control and intelligence analysis is that it is difficult to conduct interventions, either to alter the process (therapeutic interventions) or to obtain additional information (diagnostic interventions). Another distinction is that since the data is in a mostly free-form textual format, it is difficult to alarm setpoint crossings, unlike with parameter data.

- The intelligence analysis domain is a socio-technical system with many similarities to other domains studied by cognitive systems engineers. An analyst monitors a system that is complex and interconnected. Intelligence analysis is a difficult task that requires significant expertise and is performed under time pressure and with high consequences for failure.

Not surprisingly, we found that defining "data overload" was much more challenging than characterizing the main cognitive tasks and intelligence analysis domain. Although everyone in the literature agreed that data overload was an important problem that was difficult to address, the precise definition of data overload was wide-ranging. Common to most views of data overload in supervisory control domains was the notion that excessive amounts of data increased cognitive burdens for the human operator. Beyond that, however, the wide variety of design aids touted to "solve data overload" attested to the variability in definitions of the data overload problem (see Woods, Patterson, and Roth, 1998, for an extended discussion of different characterizations of the data overload problem and their associated solutions).

Given this variability in definitions of data overload, we were required to resort to "first principles" in order to come up with a definition of the data overload problem. In cognitive systems engineering, the fundamental unit of analysis is the "Cognitive Triad", which includes the demands of the work domain, the strategies of the practitioners, and the artifacts and other agents that support the cognitive processes. Since cognitive systems engineering takes the triad as the fundamental unit of analysis, we rejected definitions that isolated the data from practitioners, domain constraints, tasks, and artifacts. Therefore, we defined data overload to be a condition where a domain practitioner, supported by artifacts and other human agents, finds it extremely challenging to focus in on, assemble, and synthesize the significant subset of data for the problem context into a coherent assessment of a situation, where the subset of data is a small portion of a vast data field. The starting point for this definition was recognizing that large amounts of potentially available data stressed one kind of cognitive activity: focusing in on the relevant or interesting subset of data for the current problem context. When operators miss critical cues, prematurely close the analysis process, or are unable to assemble or integrate relevant data, this cognitive activity has broken down.

People are a competence model for this cognitive activity because people are the only known cognitive system that is able to focus in on interesting material in natural perceptual fields even though what is interesting depends on context (Woods and Watts, 1997). The ability to orient focal attention to "interesting" parts of the natural perceptual field is a fundamental competency of human perceptual systems (Rabbitt 1984; Wolfe 1992). Both visual search studies and reading comprehension studies show that people are highly skilled at directing attention to aspects of the perceptual field that are of high potential relevance given the properties of the data field and the expectations and interests of the observer. Reviewing visual search studies, Woods (1984) commented, "When observers scan a visual scene or display, they tend to look at 'informative' areas . . . informativeness, defined as some relation between the viewer and scene, is an important determinant of eye movement patterns" (p. 231, italics in original). Similarly, reviewing reading comprehension studies, Bower and Morrow (1990) wrote, "The principle . . . is that readers direct their attention to places where significant events are likely to occur. The significant events . . . are usually those that facilitate or block the goals and plans of the protagonist."

In the absence of this ability, for example in a newborn, as William James put it over a hundred years ago, "The baby assailed by eye, ear, nose, skin and entrails at once, feels it all as one great blooming, buzzing confusion" (James, 1890, I 488). The explosion in available data and the limits of current computer-based displays often leave us in the position of that baby -- seeing a "great blooming, buzzing confusion."

2.3 Diagnosis of Data Overload

Based on delving into relevant research and experience bases, we developed an explicit "diagnosis of data overload" that has proven to be a valuable synthesis about why standard ways to use technology have met with limited success across settings and what characteristics solutions to data overload will likely need to have to be effective (see Woods et al., 1998, for the complete technical report).

2.3.1 Typical "finesses" to Data Overload

We believe that the cognitive activity of focusing in on the relevant or interesting subset of the available data is a difficult task because what is interesting depends on context. What is informative is context sensitive when the meaning or interpretation of any change (or even the absence of change) is quite sensitive to some but not all the details of the current situation or past situations.

Existing techniques to address data overload often try to finesse the context sensitivity problem, that is, they avoid confronting the problem directly. Calling a technique a finesse points to a contrast. In one sense, a finesse is a positive pragmatic adaptation to difficulty. All of the finesses we identified are used to try to reduce data overload problems to manageable dimensions to allow experienced people to exhibit the fundamental human competence of extracting significance from data. However, a finesse is a limited adaptation because it represents a workaround rather than directly addressing the factors that make it difficult for people to extract meaning from data. In particular settings these finesses will be more or less brittle. Brittle techniques cope with some aspect of context sensitivity but break down quickly when they encounter more difficult cases.

Technology-centered approaches to data overload generally adopt strategies based on one or more of the following finesses because of inaccurate or oversimplified models of why data overload is a generic and difficult issue (for example, all of the following have been tried with some local success in coping with data overload in alarm systems; Woods, 1995a; Woods, 1994).

(a) scale reduction finesse – reduce available data
Scaling back the available data is an attempt to reduce the amount of stuff people have to sort through to find what is significant. The belief is that if we can reduce the size of the problem, then human abilities to find the critical data as the context changes will function adequately. Often scale reduction attempts are manifested as shifting some of the available data to more "distant" secondary displays with the assumption that these items can be called up when necessary.

This approach breaks down because of the context catch—in some contexts some of what is removed will be relevant. Data elements that appear to be less important on average can become a critical piece of evidence in a particular situation. But recognizing their relevance, finding them and integrating them in to the assessment of the situation becomes impossible if they have been excluded or pushed into the background of a virtual data world.

This finesse also breaks down because of the narrow keyhole catch—proliferating more displays hidden behind the keyhole of the CRT screen creates navigation burdens (Woods and Watts, 1997). Reducing the data available on individual displays pushes data onto more displays and increases demands for across display search and integration. This makes data available in principle, but it does not help the observer recognize or determine what would be relevant.

(b) global, static prioritization finesse -- only show what is "important"
A related finesse is to select only the "important" subset of the available data. Often, the world of data is divided into two or three "levels of importance." Domain knowledge is used to assign individual data items to one of the two or three levels. All data items identified in the highest level of "importance" would be displayed in a more salient way to users. Data elements that fall into the second or third class of less important items would be successively less salient or more distant in the virtual world of the display system and user interface.

This approach also breaks down because of the context catch—how do we know what is important without taking context into account. Context sensitivity means that it is quite difficult to assign individual elements to a place along a single, static, global priority or importance dimension. Inevitably, one is forced to make comparisons between quite disparate kinds of data and to focus on some kinds of situations and downplay others. Again, data items that are not important based on some overall criteria can be critical in particular situations.

This finesse, like the first, uses inhibitory selectivity, that is, they both, in effect, throw away data. In this case, developers will object saying that users can always call up data assigned to lower levels of importance if they feel they are relevant in a particular situation. But the problem is to help people recognize or explore what might be relevant to examine without already knowing that it is relevant. To aid this process requires one to consider perceptual organization, control of attention and anomaly recognition.

(c) intelligent agent finesse -- the machine computes what is important for you Another version of the context catch plagues this approach -- how does the machine know what is important without being able to take context into account? However, this finesse also breaks down in the face of a new catch—the clumsy automation catch. The observer now has another data source/team member to deal with when they can least afford any new tasks or any more data (Sarter et al., 1997).

The irony here is that developers believe that shifting the task to a computer somehow makes the cognitive challenges of focusing in on the relevant subset disappear. In fact, all finite cognitive processors face the same challenges, whether they are an individual, a machine agent, a human-machine ensemble, or a team of people. Just as machine diagnosis can err, we cannot expect machine agents to consistently and correctly identify all of the data that is relevant and significant in a particular context in order to bring it to the attention of the human practitioner. It always takes cognitive work to find the significance of data.

For example, attempts in the mid-80's to make machine diagnostic systems handle dynamic processes ran into a data overload problem (these diagnostic systems monitored the actual data stream from multiple sensors). The diagnostic agents deployed their full diagnostic reasoning power in pursuit of every change in the input data streams (see Woods, Pople, and Roth, 1990; Roth, Woods and Pople, 1992; Woods, 1994). As a result, they immediately bogged down, dramatically failing to handle the massive amounts of data now available (previously, people mediated for the computer by selecting "significant" findings for the computer to process). To get the diagnostic systems to cope with data overload required creating a front end layer of processing that extracted, out of all of the changes, which events were "significant" findings that required initiating a line of diagnostic reasoning. In this case determining what were significant events for diagnosis required determining what were unexpected changes (or an unexpected absence of a change) based on a model of what influences were thought to be acting on the underlying process.

(d) syntactic finesse -- use syntactic or statistical properties of text (e.g., word frequency counts) as cues to semantic content
This finesse is relied on heavily in keyword search systems, web search engines, and information visualization algorithms that utilize "similarity" metrics based on statistical properties of the text (e.g., frequency counts of different content words) to place documents in a visual space (e.g., Morse and Lewis, 1997; Wise, Thomas, Pennock, Lantrip, Pottier, Schur, and Crow, 1996). The primary limitation of this approach is that syntactic and statistical properties of text provide a weak correlate to semantics and domain content. There is rarely a simple one to one relationship between terms and concepts. It is frequently the case that one term can have multiple meanings (e.g., Ariane is both a rocket launcher and a proper name; ESA stands for the European Space Agency, Environmental Services Association, and the Executive Suite Association) and that multiple terms can refer to the same concept (e.g., the terms 'failed', 'exploded', and 'was destroyed' can be used interchangeably).

The problem is compounded by the fact that the 'relevance' metrics employed (e.g., the weighting schemes used by web search engines) are often opaque to the user. This is the lack of observability catch. The user sees the list of documents retrieved based on the query and the relevance weighting generated by the search engine. However, in many cases how the relevance weighting was generated is unclear, and the resulting document ordering does not accord well with how the user would have ordered the documents (i.e., documents that come up early with a high weighting can be less relevant than documents that come up later). This forces the user to resort to attempting to browse through the entire list. Since the generated list is often prohibitively long, it can leave the user unsure about whether important documents might be missed. Users will often prefer to browse documents ordered by metrics that do not attempt or claim to capture "relevance," such as date or source, rather than by syntactic relevance weighting because the organizing principle is observable and they know how to interpret values along those dimensions.

Attempts to place documents in a visual space based on syntactic properties are also subject to the over-interpretation catch. The spatial cues and relationships that are visible to the observer will be interpreted as meaningful even if they are incidental and not intended to be information bearing by the designer (or algorithm). For example, visualizations that attempt to represent multi-dimensional spaces (4 or more dimensions) on a two-dimensional display can create ambiguities with respect to the position of a document relative to each of the dimensions. Users may assume that two documents that are located close to each other on the display reflect a similar degree of relationship to each of the dimensions represented in the space, when in fact they are not in the same position in the multi-dimensional space – even though it looks that way on the display.

2.3.2 Constraints on Effective Solutions to Data Overload in Supervisory Control

Even before investigating intelligence analysis specifically as a domain where the data overload problem exists, we could broadly identify constraints on effective solutions to data overload in supervisory control settings from past research and experience. These constraints bound large regions of a "solution space" for where effective approaches to combat data overload are likely to exist in general but do not identify leverage points for a particular domain, as discussed in Part III.

1. All approaches to data overload involve some sense of selectivity.
However, there are different forms of selectivity: facilitation or inhibition of processing. In the former, selectivity facilitates or enhances processing of a portion of the whole. In this form of selectivity, we use positive metaphors such as a spotlight of attention or a peaked distribution of resources across the field.

In the latter, selectivity inhibits processing of non-selected areas, for example stimuli in the selected portion can pass through and go on for further processing, whereas stimuli in the non-selected portion do not go on for processing. In this form of selectivity, we use negative metaphors such as a filter or a gatekeeper.

Current research on attention suggests that we need to develop positive forms of selectivity and develop techniques that support thorough exploration of the available data. This is the case in part because observers need to remain sensitive to non-selected parts in order to shift focus fluently as circumstances change or to recover from missteps.

2. Organization precedes selectivity.
Selectivity presumes a structured field on which attention can operate, focusing on potentially interesting areas depending on context. Designers of computer technology need to define the groups/objects/events and relationships attention can select.

The default in computer systems has been to organize around elemental data units or on the units of data appropriate for computer collection, transmission, and manipulation (Flach et al., 1995). These are either too elemental, as if we saw the world in "327" variations in hue, saturation, and brightness, or too removed from the meaningful objects, events and relationships for the user's field of practice.

This finding means that effective systems for coping with data overload

- will have elaborate indexing schemes that map onto models of the structure of the content being explored, and
- will need to provide multiple perspectives to users and allow them to shift perspectives easily.

3. All techniques to cope with data overload must deal with context sensitivity. Data are informative based on relationships to other data, relationships to larger frames of reference, and relationships to the interests and expectations of the observer. Making data meaningful always requires cognitive work to put the datum of interest into the context of related data and issues.

This finding means that solutions to data overload will help practitioners put data into context. Presenting data in context shifts part of the burden to the external display rather than requiring the observer to carry out all of this cognitive work "in the head." Many techniques could support this criterion (Woods, 1995b). First, when we display a given datum, we can show it in the context of related values. Second, rather than organizing displays around pieces of data, we can organize data around meaningful issues and questions--model based displays. These are models of how data relationships map onto meaningful objects, events, and processes in the referent field of activity (Flach et al., 1995). Third, we can use the power of the computer to help extract events from the flow of elemental data. Events are temporally extended behaviors of the device or process involving some type of change in an object or set of objects. Fourth, the computer could also help observers recognize anomalies and contrasts by showing how the data departs from or conforms to the contrasting case (a departure from what is expected, from what is the plan or doctrine, from what has been typical). Since there are usually many possible contrasting cases, each defines a kind of perspective around which one views the available elemental data.

There is a prerequisite for the designer to be able to put data into context: they need to know what relationships, events, and contrasts are informative over what contexts in the field of practice.

4. Observability is more than mere data availability.

The greatest value of a picture is when it forces us to notice what we never expected to see. Tukey, 1977, p. vi

There are significant differences between the available data and the meaning or information that a person extracts from that data. Observability is the technical term that refers to the cognitive work needed to extract meaning from available data. This term captures the relationship among data, observer and context of observation that is fundamental to effective feedback.

Observability is distinct from data availability, which refers to the mere presence of data in some form in some location. For human perception, "it is not sufficient to have something in front of your eyes to see it" (O'Regan, 1992, p.475).

One example of displays with very low observability occurs on the current generation of flight decks. The flight mode annunciations are a primary indication of how automated systems are configured to fly the aircraft. These crude indications of automation activities contribute to automation surprises where the automation flies the aircraft in a way that the pilots did not anticipate. As one pilot put it, "changes can always sneak in unless you stare at it" (see Woods and Sarter, 2000, for more on this example).

Observability refers to processes involved in extracting useful information. It results from the interplay between a human user knowing when to look for what information at what point in time and a system that structures data to support attentional guidance (see Rasmussen, 1985; Sarter, Woods and Billings, 1997). The critical test of observability is when the display suite helps practitioners notice more than what they were specifically looking for or expecting. If a display only shows us what we expect to see or ask for, then it is merely making data available.

5. To cope with data overload, ultimately, will require the design of conceptual spaces.
One builds a conceptual space by depicting relationships in a frame of reference (Woods, 1995b; Rasmussen et al., 1994). The search to solve data overload begins with the search for frames of reference that capture meaningful relationships for that field of practice. A frame of reference is a fundamental property of a space and what makes a space or map special from the point of view of representation. With a frame of reference comes the potential for concepts of neighborhood, near/far, sense of place, and a frame for structuring relations between entities. A frame of reference is a prerequisite for depicting relations rather than simply making data available.

Almost always there are multiple frames of reference that apply. Each frame of reference is like one perspective from which one views or extracts meaning from data. Part of designing a conceptual space is discovering the multiple potentially relevant frames of references and finding ways to integrate and couple these multiple frames.

We now compactly summarize our diagnosis of the data overload problem:

- What is the definition of data overload? Data overload is a condition where a domain practitioner, supported by artifacts and other human agents, finds it extremely challenging to focus in on, assemble, and synthesize the significant subset of data for the problem context into a coherent assessment of a situation, where the subset of data is a small portion of a vast data field.
- Why is data overload so difficult to address? Context sensitivity – meaning lies, not in data, but in relationships of data to interests and expectations.
- Why have new waves of technology exacerbated, rather than resolved, data overload? When they ignore or finesse context sensitivity.
- How are people able to cope with data overload? People are able to shift focus of attention fluently as circumstances change and re-orient to potentially interesting new stimuli.

PART III. CALIBRATING RESEARCH BASE TO INTELLIGENCE ANALYSIS

Although much can be gained by using research bases to jumpstart a project, not everything translates from the research bases to a specific setting like intelligence analysis (Figure 6). An important step before committing to a particular research or design direction is to identify the predominant themes and challenges in a setting given the specific tasks, support tools, distribution of cognition, practitioners, and organizational factors. In a project that cross-fertilizes research and design efforts, we try to explicitly identify the abstract problem of which a particular problem is an instantiation while also targeting the research/design activities at domain-specific leverage points. Leverage points depend on domain-specific attributes such as the most prominent aspects of the data overload problem, the format and representation of data, the unique cognitive challenges and strategies that are employed to address those challenges in a work setting, the nature of tools available to practitioners, the risks of tradeoff options when goals conflict, and the consequences of particular types of failures.

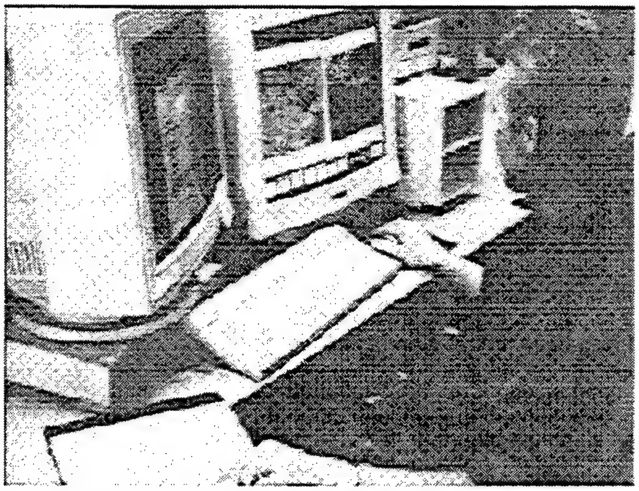


Figure 6. Computer-Supported Intelligence Analysis

To this end, a target situation was simulated. The target situation was inferential analysis conducted by experienced analysts under data overload, on tight deadlines, and outside their immediate bases of expertise. The study was designed to simulate this target situation:

1. ten professional intelligence analysts, ranging from 7 to 30 years of analytic experience, representing diverse areas of expertise that were related to portions of the simulated task,
2. analyzing a face valid task that they had not previously analyzed and was not in the immediate base of expertise: the cause and impacts of the June 4, 1996, Ariane 501 rocket launch failure on the Ariane 5 rocket's maiden flight,

3. given 2000 text documents in a mostly "on topic" database generated by representative searches in Lexus Nexus™ and DIALOG™ by the investigators and a professional search intermediary from the intelligence agency,
4. in 3-4 hour sessions, and
5. using a "baseline" toolset that supported keyword queries, browsing articles by dates and titles sorted by relevance or date, and cutting and pasting selected portions of documents to a text editor.

3.1 Scenario Design

Because intelligence analysis was a fairly new domain for the investigators, it was judged that there was an insufficient research base to carefully craft a new scenario that instantiated the demands in intelligence analysis. Nevertheless, several models from the research bases were leveraged in selecting a scenario for the simulated task. Primarily, the scenario that was selected involved an organizational investigation of a costly accident in a complex system. Much was known about this kind of situation from previous investigations of accidents in health care, aviation, and nuclear power. Therefore, the Ariane 501 rocket launch failure was selected as the task to be simulated by experienced analysts.

The Ariane 501 accident scenario was selected as the analysis task to be performed by study participants under the conditions of data overload and a short deadline of several hours. The maiden launch on June 4, 1996, of the Ariane 5 vehicle ended in a complete loss of the rocket booster and the scientific payload, four Cluster satellites, when it exploded 30 seconds after liftoff. The Ariane 5 rocket was a new European rocket design by Arianespace that was intended to eventually replace the successful Ariane 4 rocket. The Ariane 5 rocket was designed to be larger, more powerful, and to carry multiple payloads. The Ariane 501 accident was significant in how it departed from typical launch failures. First, the explosion was due to a design problem in the software rather than the more classic mechanical failure – there was numerical overflow in an unprotected horizontal velocity variable in the embedded software that was re-used from the Ariane 4, which was a slower rocket. Additionally, it was the first launch of a new rocket design, which raised concern about the viability of the new design. Overall, however, launch failures were relatively common in the industry, and first launches in particular were prone to fail, so the reputation of the Ariane program was not greatly damaged.

From interviews prior to the study, it was confirmed that the Ariane 501 scenario captured critical aspects necessary for high face validity for the study participants. First, the scenario was challenging to analyze in a short time, with opportunities for the study participants to make inaccurate statements based on misleading and inaccurate information in the database provided to them. Second, the scenario required technical knowledge about the engineering design of aerospace vehicles, which was prototypical of tasks performed by analysts at the agency. Third, although all of the study participants had some relevant experience that helped them to perform the analysis, none of the participants had been directly monitoring the particular country or technologies involved in the scenario. Fourth, unclassified sources such as Aviation Week and Space Technology were available for the scenario that closely paralleled classified sources in reporting style, depth of the analyses, and technical knowledge of the reporters.

The Ariane 501 scenario had the additional benefit of involving an organizational investigation of an accident, which allowed us to leverage conceptual frameworks in designing the simulated task. For example, the dates for documents in the electronic database ranged from 1994 until 1999 (Figure 7). These dates were selected so that the database included distractors prior to the accident, the Ariane 501 accident, the Inquiry Board Report detailing the findings of the accident investigation, and the next landmark event after the accident, the Ariane 502 launch. The naturally emerging structure of reports in the database mirrored structures from accident investigations in other complex settings. The initial flurry of reports about the accident tended to be sensationalistic, included quotes from eyewitnesses about the causes and immediate reactions to the accident from affected parties, and contained details not available in later reports, some of which later turned out to be inaccurate. These early reports emphasized contributors to the accident that were closest in time and space (e.g., decision by ground operator to blow up the rocket). The second main flurry of reports summarized the findings of the Inquiry Board about the causes of the accident. Intermittently after this flurry, comprehensive, in-depth analyses of the accident and long-term impacts could be found. Reports at this time tended to have less diversity in the descriptions about the causes and impacts of the accident and contained fewer details. These later reports included contributors that were farther in space and time from the accident -- limitations with the design and testing of the rocket and the organizational context for the rocket design. Finally, another small flurry of reports was seen immediately following the next attempted launch of the Ariane 5 rocket, which was the next landmark event after the accident. These reports briefly mentioned the Ariane 501 accident and provided updates on several themes.

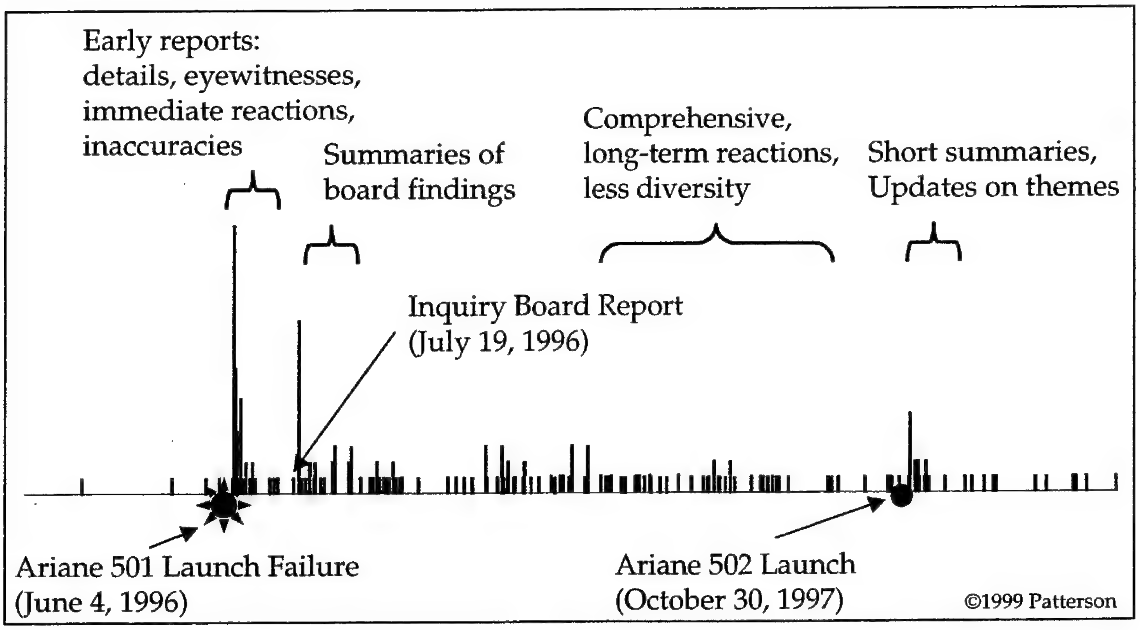
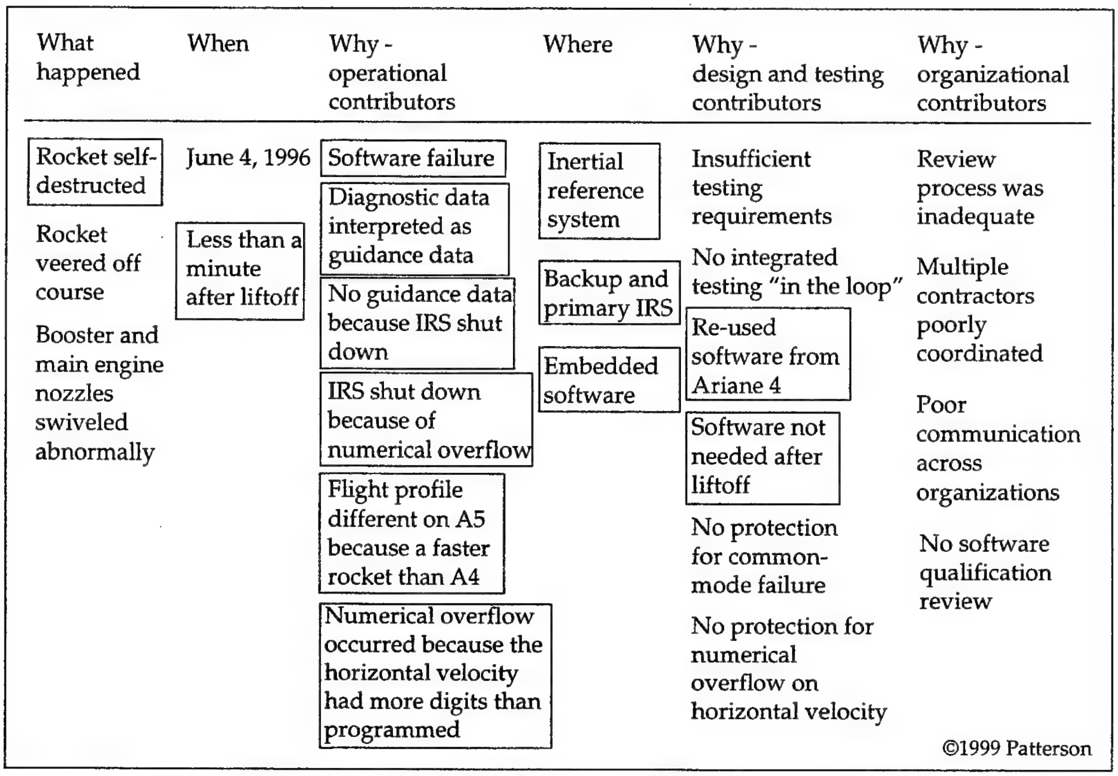


Figure 7. Report Database Included Typical Reactions to Accident

In addition, discrepancies in the data set provided to the participants followed patterns seen in other accident investigations in complex systems (boxed items in Figure 8 had inaccurate information in the database about that item). There were inaccuracies in early reports about the causes of the accident because all of the data was not yet available. For example, it was reported that ground controllers blew up the rocket when it actually had self-destructed because the initial reports were based on seeing the ground controller push the destruct button, although by then the rocket had already self-destructed. In addition, there were inaccuracies that stemmed from a lack of in-depth, technical knowledge on the part of the reporters, such as why there was a numerical overflow.



What happened	When	Why - operational contributors	Where	Why - design and testing contributors	Why - organizational contributors
Rocket self-destructed Rocket veered off course Booster and main engine nozzles swiveled abnormally	June 4, 1996 Less than a minute after liftoff	Software failure Diagnostic data interpreted as guidance data No guidance data because IRS shut down IRS shut down because of numerical overflow Flight profile different on A5 because a faster rocket than A4 Numerical overflow occurred because the horizontal velocity had more digits than programmed	Inertial reference system Backup and primary IRS Embedded software	Insufficient testing requirements No integrated testing "in the loop" Re-used software from Ariane 4 Software not needed after liftoff No protection for common-mode failure No protection for numerical overflow on horizontal velocity	Review process was inadequate Multiple contractors poorly coordinated Poor communication across organizations No software qualification review

©1999 Patterson

Figure 8. Discrepancies in the Causes of the Ariane 501 Failure

Finally, information about impacts of the accident on the Ariane 4 rocket program, Cluster scientific program, and the next launch in the Ariane 5 rocket program, 502, came in over time, causing information from different points in time to conflict (see Figure 9). For example, the original predictions of the second launch of the Ariane 5 vehicle (502) of September 1996 were overly optimistic, and predictions gradually became nearer to the actual date of October 30, 1997. As is expected following surprising accidents, predictions about the impacts radically changed over time. For example, immediately following the 501 accident, it was reported that the Cluster scientific program would be shut down because of the uninsured loss of the $500 million scientific satellites. A month later, it was reported that one of the four satellites might be rebuilt. Two months later, it was reported that either one or four satellites would be rebuilt. Seven months later, it was reported that all four of the satellites would be rebuilt.

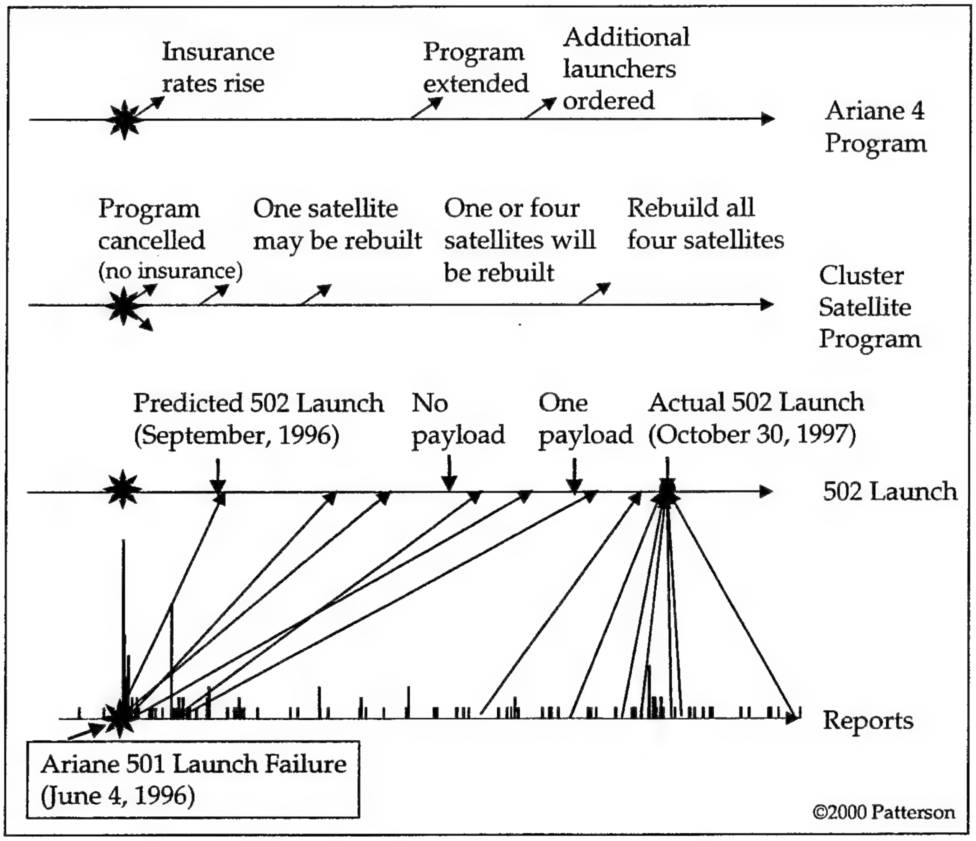

Insurance rates rise
Program extended
Additional launchers ordered
Ariane 4 Program
Program cancelled (no insurance)
One satellite may be rebuilt
One or four satellites will be rebuilt
Rebuild all four satellites
Cluster Satellite Program
Predicted 502 Launch (September, 1996)
No payload
One payload
Actual 502 Launch (October 30, 1997)
502 Launch
Reports
Ariane 501 Launch Failure (June 4, 1996)
©2000 Patterson

Figure 9. Changing Assessments of Impacts of the Ariane 501 Failure

The database provided to the study participants contained enough information to conduct a comprehensive analysis of the causes and impacts of the Ariane 501 accident. There were approximately 2000 unclassified text documents. The majority (~60%) of the documents were "on target" in that they contained information about the causes and impacts of the accident. Some of the documents (~35%) contained information that helped to provide context, such as information about other rocket launch failures, but were not directly relevant to the task. As would be expected by intelligence analysts from searches of their organizational databases, in contrast to keyword searches on the World Wide Web, only a small portion contained completely irrelevant information (~5%), such as articles about women named Ariane. Nine documents in the database were identified as particularly high quality, classified as "high profit" documents, by the investigators. The high profit categorization was based on both high topicality and utility, which are often used in relevance definitions in the information retrieval literature (see Mizarro, 1997, for an overview of the factors in relevance definitions; cf. Blair and Maron, 1985, for their distinctions between vital, relevant, partially relevant, and not relevant documents in legal analysis). An example of a high profit document was the Inquiry Board Report from the European Space Agency.

We would like to note that we devoted a significant amount of energy to the selection of the scenario to be used in the study. We feel that this is indicative of an approach that values complementarity between research and design. The scenario selection is critical if we are to learn the cognitive challenges in a domain. It is critical if we want to identify leverage points for design directions to combat data overload. The crafting of the problem to be solved is critical given the framing of cognitive systems engineering research as computer-supported problem solving by practitioners with expertise.

Before selecting the Ariane 501 scenario, we considered other scenarios. First, we considered using a real-world analysis case. This was not possible because of restrictions on classified information. Even if we had selected an analysis and only used unclassified information in the study, the result of the analysis might have become classified because of the additional information and value that an intelligence analyst performing the analysis would bring. Then we considered using a terrorism training case involving the countries Aland, Beeland, and Ceeland. We decided against using this scenario because we felt that it lacked sufficient distractors in the data, that it would be important to have a more challenging scenario in order to better understand the nature of analytic expertise, and that adding aspects that were important to our conceptual focus such as data overload conditions would be difficult to generate without more domain knowledge. We also considered using the Zairean civil war as a case based largely on New York Times articles. In interviews with analysts, we discovered that this case had low face validity because there was no technological component. The weapons and transportation available in Zaire are not as technologically advanced as would normally be the case for most analysis tasks. In addition, we learned that sources like New York Times lacked face validity because they assume a largely uninformed audience. Essentially, every article about the conflict contained mostly the same information with only minor updates because the assumption was that the audience knew little to nothing about the situation. Based on this feedback, we considered crafting a scenario in collaboration with an expert analyst that contained challenges similar to the Zairean conflict but with higher face validity. We did not further pursue this strategy because of the difficulty of crafting a scenario in a domain that was new to us, even in partnership with intelligence analysts. Finally, we identified the competition between Airbus and Boeing in China as a potential scenario using unclassified but technical articles, such as from Aviation Week and Space Technology. We believe that this scenario could form the basis for a useful case, but we opted to use Ariane 501 instead because it would be easier to have questions with varying levels of difficulty because there was a dominant, distinct event on a particular date in case we encountered a floor or ceiling effect during the pilot study. Therefore, although we selected rather than crafted a scenario in this instance, it is important to note that the same considerations about mapping a target to a test situation, the level of difficulty in the scenario, face validity, and the ability to compare performance against a normative standard applied in the selection criteria.

3.2 Analysis Methodology

The study participants were asked to think aloud during the simulated analysis task and provide a verbal briefing in response to the written question: "In 1996, the European Space Agency lost a satellite during the first qualification launch of a new rocket design. Give a short briefing about the basic facts of the accident: when it was, why it occurred, and what the immediate impacts were?" Two investigators (EP, ER) directly observed this process for all of the study participants, which was also audio and videotaped. The investigators noted during the session what articles were opened by the participants. The investigators also electronically saved the queries, the documents that were returned by the queries, the documents that were marked with the marking function in the software, the workspace configuration of the screen, and snapshots of the electronic notes generated during the session.

A collection of four protocols which emphasized different aspects of the simulated task were generated: the search strategies, the selection and interpretations of documents, the strategies for resolving conflicts in the data, and the construction of the verbal briefing. An excerpt of the protocol focusing on document selection and interpretation is provided in Figure 10. This protocol was generated by a single investigator and then verified and expanded by another investigator. Differing interpretations of the data were identified in this fashion and resolved through additional searches for converging evidence and debate.

Article #	Query	Name and source info	Why selected	Important?	Notes
1380	1	ARIANE 5 EXPLOSION CAUSED BY FAULTY SOFTWARE; SATELLITE NEWS	wants to work backwards so wants a late article		faulty software
1274	1	NEW CLUES TO ARIANE-5 FAILURE; DEFENSE DAILY	title and looking for date of event		June 4, 1996 (limits query results to after June 1 since event is June 4)
253	1A	STRIDE: FIRING TESTS OF NEW H IIA ROCKET ENGINE COMPLETED	time of article close to event		of no interest -- recognizes the HIIA rocket engine is from Japan
1855	1A	European space rocket explodes: Work continues with 14 similar models; Ottawa Citizen		Cuts and pastes	5 km from launch site 40 seconds 14 rockets on production line – if fault is not generic, the program won't suffer too much (software would classify as not generic according to him)
1223	1A	False computer command blamed in Ariane V failure; Aerospace Daily	6-6-96 date, also title	Cuts and pastes, marks, says good article	computer command Aerospace Daily as a good source says article is "remarkably good" and takes a while reading it June 6 knew false signal and looking closer at it Says what causes were eliminated

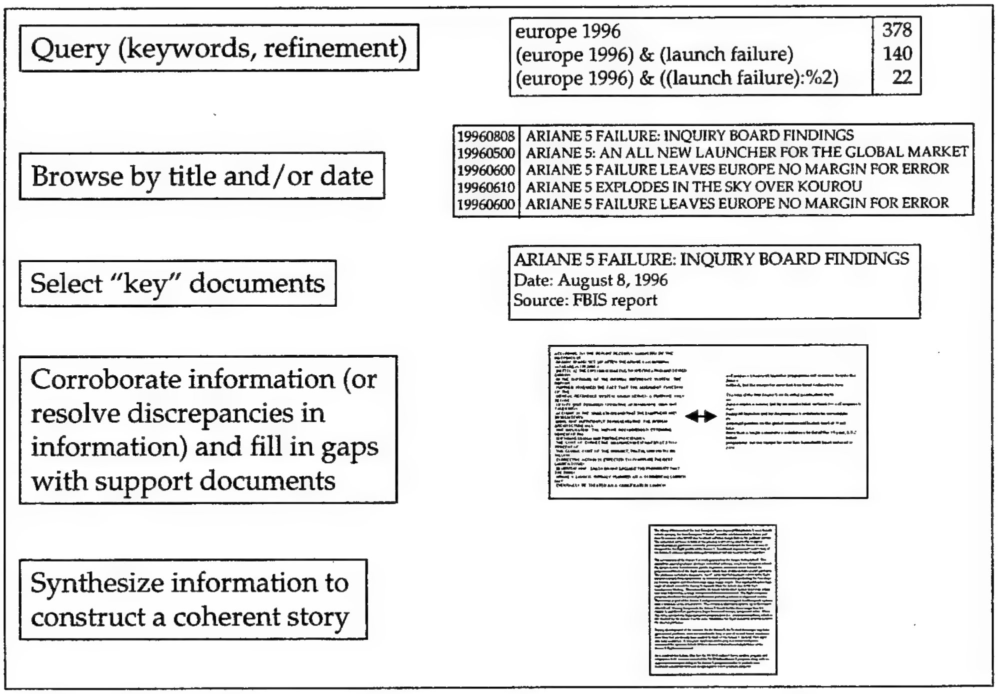
Figure 10. An Excerpt of Participant 5's Article Trace Protocol

The data analysis was an iterative, discovery-oriented process. As the base protocols were generated, potential areas were noted for more detailed investigation. At the same time that the analytic processes were being traced, the "products" of the analysis (i.e., the verbal briefings) were investigated. The verbal briefings were transcribed and items were coded as not mentioned, accurate, vague, and inaccurate. The process that an individual participant followed that arrived at the inaccurate statement was analyzed and emerging patterns used to identify the cognitive challenges that led to inaccurate statements across participants.

In summary, the analysis process involved bottom-up searching for patterns combined with top-down conceptually driven investigations (see Woods, 1993, for a description of the process tracing methodology used in the data analysis). The base protocols served as a detailed account of the process from the perspective of different conceptual frameworks, including strategies to cope with data overload in supervisory control, information retrieval strategies, and resolving data interactions in abductive inference. The protocols were used to identify patterns on particular themes. These patterns were then represented across participants in ways that highlighted similarities and differences along relevant dimensions.

3.3 Study Findings

The analysis process employed by all the study participants generally followed the pattern shown in Figure 11. Reports were selected from the database through the refinement of keyword queries and by browsing the returned reports by title or date. A small number of their sampled reports were heavily relied upon, which we refer to as "key" documents. The key documents made up the skeleton of the analysis product. Excerpts from supporting documents were then used to corroborate some of the information and fill in details. Conflicts in the data were flagged and judgments about which data to include in the developing story were revisited as new information on the topic was discovered. When the study participants felt ready, they organized their notes and generated a coherent story.


Query (keywords, refinement)
Browse by title and/or date
Select "key" documents
Corroborate information (or resolve discrepancies in information) and fill in gaps with support documents
Synthesize information to construct a coherent story

europe 1996	378
(europe 1996) & (launch failure)	140
(europe 1996) & ((launch failure):%2)	22

19960808	ARIANE 5 FAILURE: INQUIRY BOARD FINDINGS
19960500	ARIANE 5: AN ALL NEW LAUNCHER FOR THE GLOBAL MARKET
19960600	ARIANE 5 FAILURE LEAVES EUROPE NO MARGIN FOR ERROR
19960610	ARIANE 5 EXPLODES IN THE SKY OVER KOUROU
19960600	ARIANE 5 FAILURE LEAVES EUROPE NO MARGIN FOR ERROR

ARIANE 5 FAILURE: INQUIRY BOARD FINDINGS
Date: August 8, 1996
Source: FBIS report

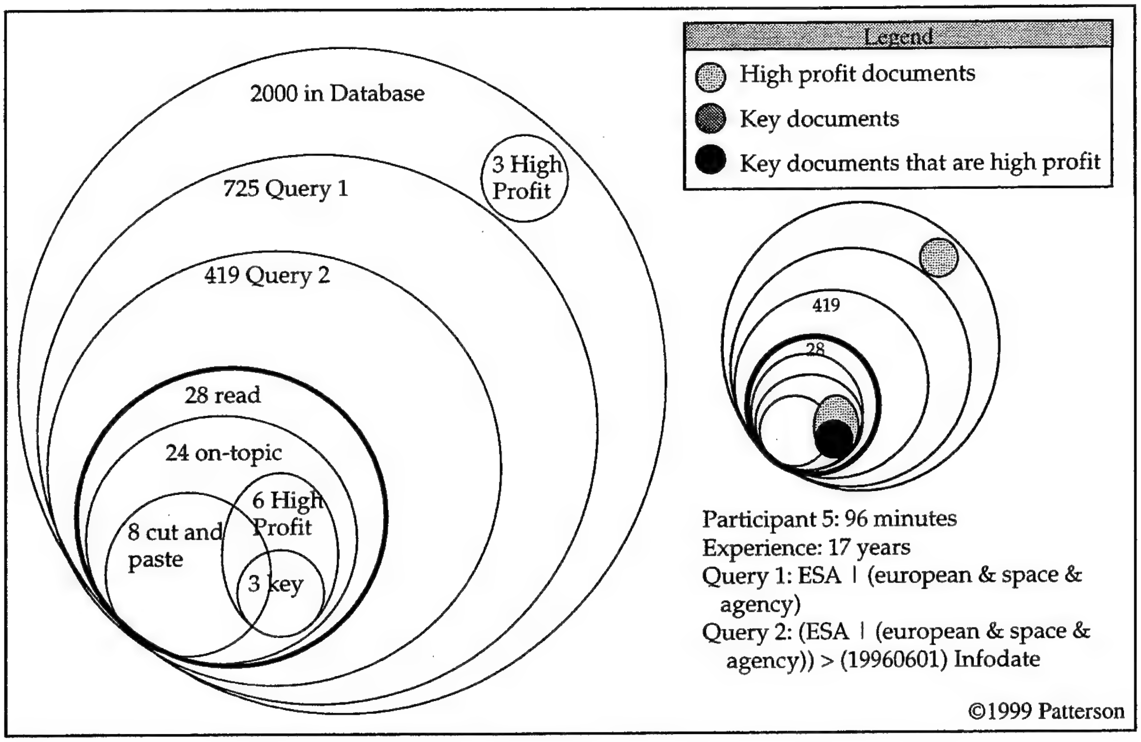
Figure 11. Typical Analysis Process

3.3.1 Search Strategy: Sampling Documents by Narrowing In

In inferential analysis under data overload in baseline electronic environments with textual databases, information is effectively sampled, generally through querying and browsing. In our study, participants were observed to begin the analysis process by making queries with standard inputs such as keywords and date limits. If a returned set of documents was judged to be too large, the search was narrowed rather than starting with a new set of search terms. Typical narrowing strategies included adding a keyword, limiting to a date range, or enforcing a proximity requirement on a set of keywords. The search was then further narrowed through the process of browsing by summary information about a document, typically dates and titles. Then documents were opened by double-clicking on a report title.

A subset of the opened documents was judged to be relevant to the analysis. Of this set of documents, a small number were used as the basis for the analysis, which we refer to as "key" documents. For this study, the definition of what documents were treated as "keys" was based on converging behavioral and verbal data from the process traces. The key documents were associated with verbalizations such as "Here we go!" or "That's a good one!" In addition, the participants were often observed to spend a longer time reading them than other documents, copy much of the document to their electronic notes, and/or use the marking function in the database software to highlight the title in the browsing window. Convergingly, the phrases used in the verbal briefings provided evidence for what documents were heavily relied upon in the analysis process.

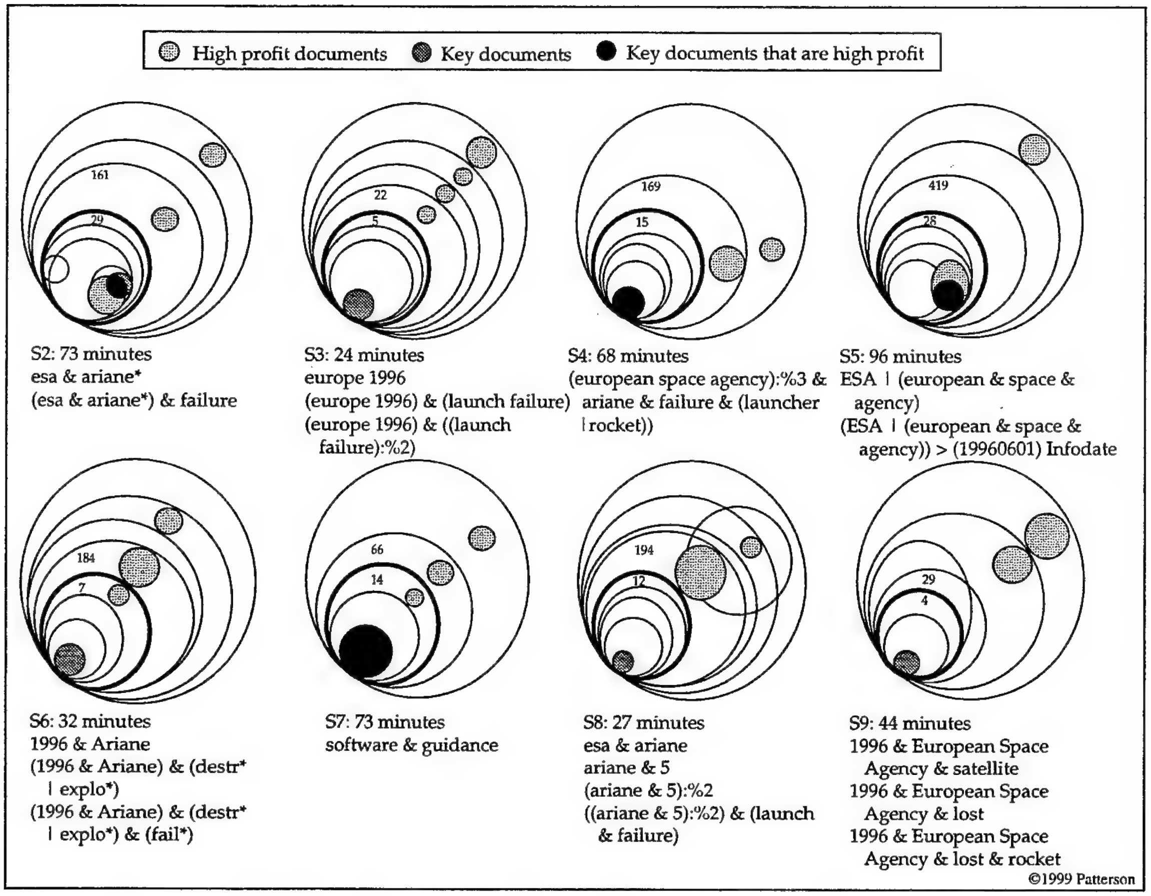
To illustrate this process, consider the information sampling process employed by study participant 5 during the analysis (Figure 12). The participant started with a Boolean keyword search (esa OR (european AND space AND agency). This search returned 725 hits, so he narrowed the search to documents published after June 1, 1996 after determining that the date of the accident was June 4, 1996 from scanning three articles. 419 documents remained after this narrowing criteria, which became his "home query" in that he did no more keyword searches. Twenty-eight documents were opened during the analysis, 24 of which were on-topic, or relevant to the analysis. Six of the documents that he opened were "high-profit" in that they were judged by the investigators to be highly informative documents. The other three high-profit documents were available in the database but were not returned by either query. The participant cut and pasted portions of eight documents along with references into a word processing file and used a marking function in the software to highlight two documents, one because he stated that it was a remarkably good article and one to mark in case he needed to refer back to it later in the analysis for further information. Three articles were identified as his "key" documents – 1) document 1223 because he remarked that it was "remarkably good" and spent a long time reading it, 2) document 1301 because he spent a long time reading it and made many verbalizations about details of the accident while reading it and said after reading it that now he had a good idea of what had happened, and 3) document 1882 because he said that it was "a definite keeper," that it was like briefings by professional analysts in its quality, spent a long time reading it, cut and pasted the most text from it, and made many verbalizations while reading it. All three of his key documents were high profit documents.



Participant 5: 96 minutes
Experience: 17 years
Query 1: ESA | (european & space & agency)
Query 2: (ESA | (european & space & agency)) > (19960601) Infodate

Figure 12. Information Sampling Process Employed by Study Participant 5

The information sampling strategy for study participant 5 was essentially one of continually narrowing in. An initial query was refined to reach a document set that was judged manageable based on the number of hits. A small subset of these documents was then heavily relied upon in generating the analysis product. Looking at the searching processes for all of the study participants (Figure 13), this process was representative. All of the participants narrowed their queries to a number that they judged to be manageable (22 – 419 documents) from which they opened documents based on a view of the dates and titles (4 – 29 documents). They then relied heavily on a subset of these documents (1-4 documents) for their verbal briefings.



High profit documents
Key documents
Key documents that are high profit
161
29
22
5
169
15
419
28
S2: 73 minutes
esa & ariane*
(esa & ariane*) & failure
S3: 24 minutes
europe 1996
(europe 1996) & (launch failure)
(europe 1996) & ((launch failure):%2)
S4: 68 minutes
(european space agency):%3 & ariane & failure & (launcher | rocket))
S5: 96 minutes
ESA | (european & space & agency)
(ESA | (european & space & agency)) > (19960601) Infodate
184
7
66
14
194
12
29
4
S6: 32 minutes
1996 & Ariane
(1996 & Ariane) & (destr* | explo*)
(1996 & Ariane) & (destr* | explo*) & (fail*)
S7: 73 minutes
software & guidance
S8: 27 minutes
esa & ariane
ariane & 5
(ariane & 5):%2
((ariane & 5):%2) & (launch & failure)
S9: 44 minutes
1996 & European Space Agency & satellite
1996 & European Space Agency & lost
1996 & European Space Agency & lost & rocket
©1999 Patterson

Figure 13. Searching Process Employed by all Study Participants

This pattern suggests that, under data overload conditions, narrowing in on a small subset of information is a commonly used coping strategy. Others have observed this propensity to narrow returned sets based on the number of hits almost indiscriminately when the data sets are large (Blair, 1980 observed this pattern with users of indexed databases and explained the pattern as a result of overestimating the probability of conjunctive sets; Olsen, Sochats, and Williams, 1998 discuss the overuse of adding keyword terms to narrow document sets). Although effective in making the amount of data to be browsed manageable, this coping strategy leaves analysts vulnerable to missing critical information, such as the high profit documents not opened by the study participants.

The narrowing strategies employed by the participants are relatively primitive compared to tactics described in the information retrieval literature (see Bates, 1979 for search tactics to narrow the number of documents that are returned by a query). The emphasis appears to be on quickly getting to a number of documents that can be browsed rather than seeking a high quality, precise, or

exhaustive set of information. For example, the participants did not use orthogonal facets to narrow the number of returned hits. This strategy would involve combining synonyms with an "OR" command crossed by orthogonal facets with an "AND" command. Instead, some of the terms that were used to narrow the search were synonyms, such as when fail* was ANDed with a query combination that already included (destr* OR explo*) by participant 6.
The finding that the study participants used relatively primitive search strategies is not surprising in the context of the growing information retrieval literature on other domain expert end-users who conduct their own searches but are not search experts (e.g., legal analysts, Blair and Maron, 1985). Across a number of studies, there is converging evidence that although domain experts can quickly learn to conduct simple searches, many never learn to employ more sophisticated search techniques.

One caution in determining implications for this finding is that this does not necessarily imply that all intelligence analysts should use professional search intermediaries to perform their searches. It is a consistent finding in information retrieval studies that both domain knowledge and search expertise are important in seeking information, and that one is not significantly more important than the other (Saracevic, Kantor, Chamis, and Trivison, 1988). Also, these two sources of knowledge are only partially decomposable, and may in fact interact in important ways (Shute and Smith, 1992).

It is not surprising, given the type of computer support that was provided to the participants, that all of the participants missed high profit documents without being aware of it (cf., Blair and Maron's 1985 landmark study of legal analysts who were poorly calibrated to the amount of relevant information that they were missing from searching an electronic database). Samples that were returned by the keyword searches were essentially opaque in terms of how they related to what was available, such as what high profit documents were left out of the query results. Then documents were sampled based on a view of the dates and titles, which were also weak indicators of whether or not documents were high profit, as can be seen in Table 1 where high profit documents and documents that were particularly poor quality are indistinguishable. The first "low profit" article was a translated description of an article originally published in Italy that contained inaccuracies about the details of the cause of the software failure. The second article was a one-paragraph abstract and so contained very little information. The third article contained significant inaccuracies because it was published soon after the event occurred.

Table 1. Dates and Titles of Low and High Profit Articles

"Low-profit" articles	**"High-profit" articles**
Europe: Causes of Ariane 5 Failure (July 5, 1996)	Software design flaw destroyed Ariane V; next flight in 1997 (July 24, 1996)
Ariane 5 Failure: Inquiry Board Findings (July 25, 1996)	Board Faults Ariane 5 Software (July 29, 1996)
False computer command blamed in Ariane V failure (June 6, 1996)	Ariane 5 loss avoidable with complete testing (September 16, 1996)

Note that during the process of searching for information, some study participants verbalized that perhaps they should conduct new searches for specific information, but did not. In addition, comments made by some of the study participants indicated that they did not know what was available in the database and how their queries related to what was available, which made them uncomfortable. In spite of these statements, the study participants appeared reluctant to leave the working area that the home query window represented. The participants developed a familiarity with the titles and dates of the documents returned by the query, the participant had often sorted the documents by date, the windows had been resized and placed in a dedicated place on the screen, and some of the documents had been marked for various reasons.

3.3.2 Basing Analyses on High Profit Documents

Looking more closely at the process traces in Figure 13, the black circles represent when the key documents were also high profit documents, or in other words, when the documents that were heavily relied upon were the best documents available in the database. Comparing the four participants that used some high profit documents as key documents vs. the four that did not, there are some interesting differences between the two groups (Tables 2 and 3). The participants that used high profit documents as key documents spent more time during the analysis, read more documents, and read more of the high profit documents.

Table 2. Participants That Used High Profit Documents as Key vs. Not

Participants whose key documents were not high profit documents

Participant	Experience (years)	Time (mins.)	Final query (no. hits)	Documents (no. read)	High profit docs (no. read)
3	7	24	22	5	0
6	8	32	184	7	2
8	11	27	194	12	0
9	18	44	29	4	0
Average:	11	32*	107	7*	0.5*

Participants whose key documents were high profit documents

Participant	Experience (years)	Time (mins.)	Final query (no. hits)	Documents (no. read)	High profit docs (no. read)
2	8	73	161	29	3
4	8	68	169	15	2
5	17	96	419	28	2
7	9	73	66	14	5
Average:	10.5	78*	204	22*	3*

* significant difference using Wilcoxon-Mann-Whitney Non-Parametric test (Siegel and Castellan, 1988)

Table 3. Comparison of Querying and Browsing Breadth

Participants whose key documents were not high profit documents

	Final "Home" Query	No. of Hits in Query	No. of High Profit Hits in Query	Percent of Query Docs that are High Profit	No. of Documents Read	No. of High Profit Documents Opened	Percent of "Key" Docs that are High Profit
3	(europe 1996) & ((launch failure):%2)	22	1	5%	5	0/9	0% (0/1)
6	(1996 & Ariane) & (destr* | explo*) & (fail*)	184	7	4%	7	2/9	0% (0/3)
8	((ariane & 5):%2) & (launch & failure)	194	8	4%	12	0/9	0% (0/1)
9	1996 & European Space Agency & satellite & lost & rocket	29	0	0%	4	0/9	0% (0/1)
	Average:	107	4	3%	7*	0.5/9*	0%

Participants whose key documents were high profit documents

	Final "Home" Query	No. of Hits in Query	No. of High Profit Hits in Query	Percent of Query Docs that are High Profit	No. of Documents Read	No. of High Profit Documents Opened	Percent of "Key" Docs that are High Profit
2	(esa & ariane*) & (failure)	161	6	4%	29	3/9	50% (1/2)
4	(european space agency):%3 & ariane & failure & (launcher | rocket)	169	7	4%	15	2/9	100% (2/2)
5	(ESA | (european & space & agency)) > (19960601) Infodate	419	7	2%	28	2/9	33% (1/3)
7	Software & guidance	66	7	11%	14	5/9	100% (4/4)
	Average:	204	7	5%	22*	3/9*	71%

* significant difference using Wilcoxon-Mann-Whitney Non-Parametric test (Siegel and Castellan, 1988)

We believe that the best explanation for the differences between these two groups is that the participants who found the high profit documents were more "persistent" in that they took longer and read more documents. It follows that they were therefore more likely to find the high profit documents. There could be alternative explanations for the differences between these two groups. It is generally recognized in the information retrieval literature that both search and domain expertise is important in information seeking. Therefore, it is possible that the group of analysts that relied on the high profit documents used more effective search strategies to find the documents. Similarly, it is possible that the more experienced professional analysts had developed strategies that helped them to perceive high profit documents, or that domain- or scenario-related expertise would make it easier for them to recognize high profit documents. We investigated nine potential hypotheses relating to these possibilities and found little support for these alternative explanations (Patterson, Woods, and Roth, 1999).

3.3.3 Impact of Basing Analyses on High Profit Documents

An important question to answer is whether the study participants who used the high profit documents as key documents in their analyses performed better than those that did not. Although analysts in prior interviews had described that they considered it critically important to have high-quality documents, it is possible that they had developed expert strategies that allowed them to use converging information from lower quality sources in such a way as to perform well despite having to rely lower quality information.

To this end, the study participants' verbal briefings were coded on 20 topic items from the Ariane 501 case as accurate, vague, inaccurate, or no information (Table 4)[2]. It appears that there might in fact be differences in performance between the participants who relied upon the high profit documents and the participants who did not. As would be expected, the participants who relied on high profit documents in their analysis had fewer inaccurate statements in their verbal briefings than the other participants who had some of their key documents be high profit documents (1 vs. 6, p = 0.03). Note that this difference is not explained by one group of participants having more thorough analyses, thereby increasing the likelihood of inaccurate statements, because there were no significant differences between the two groups in the overall number of items included in the briefings. Also, years of analytic experience is not significantly different between the groups (11 years vs. 10.5 years).

[2] Intercoder reliability by two simultaneous coders was 84% for the eight study participants. The discrepancies were resolved by discussion and both coders agreed to the final codes.

Table 4. Summary of Types of Statements in Verbal Briefings

Participants whose key documents were not high profit documents

Participant	Accurate	Vague	Inaccurate	Nothing
3	5	2	2	11
6	11	1	3	5
8	9	0	0	11
9	5	3	1	11
Average:	7.5	1.5	1.5*	9.5

Participants whose key documents were high profit documents

Participant	Accurate	Vague	Inaccurate	Nothing
2	5	2	0	13
4	11	2	0	7
5	12	3	0	5
7	8	1	0	11
Average:	11	2	0*	6.75

* significant difference using Wilcoxon-Mann-Whitney Non-Parametric test (Siegel and Castellan, 1988)

3.3.4 Sources of Inaccurate Statements

Two main conceptual frameworks were used to look for patterns in the analytic processes. The first framework was information sampling strategies, generally referred to as search tactics in the information retrieval literature. The second framework was evidence interactions in abductive inference (Josephson and Josephson, 1994), which is inference to the best explanation. Diagnosis is an example of a well-known abductive inference process, where a diagnostic reasoner selects an explanatory hypothesis to explain observed symptoms. The abductive process involves observing deviations from a nominal state, proposing explanatory hypotheses to account for the deviations, and selecting the "best" or most warranted explanation from the hypothesis set.

Determining the cause of the Ariane 501 accident could be characterized as an abductive inference task. There is anomalous data that could be explained by several hypotheses (Figure 14). For example, the observation that the rocket swiveled abnormally could have been due to poor guidance data, a mechanical failure, or a software failure. The main observation that pointed to a software failure hypothesis rather than other hypotheses was that both the primary and backup Inertial Reference Systems (IRS) shut down simultaneously. Although this finding made the software failure the most plausible explanation, there was an additional finding that was not covered by this hypothesis -- unexpected roll torque during ascent. The full set of observations was explained by the combination of two hypotheses – a software failure and an unrelated mechanical problem.

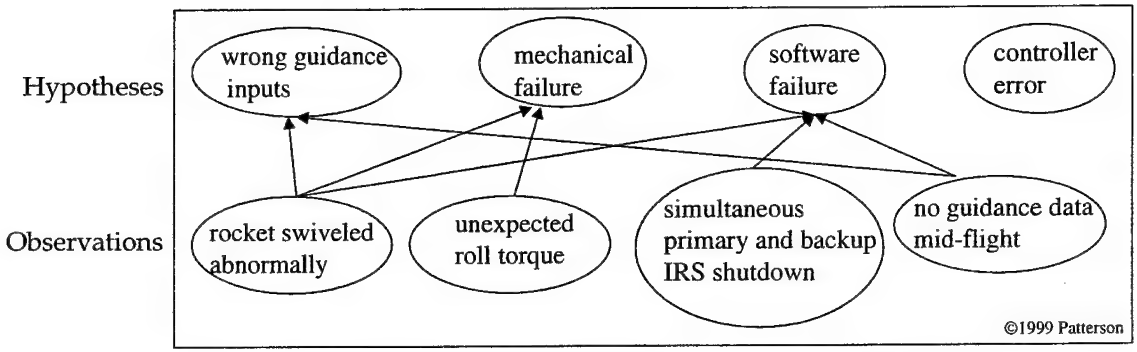


Figure 14. Hypothesis Space in Ariane 501 Scenario

During the data analysis, we were surprised to discover that there was remarkably little evidence from the think-aloud protocols and decisions regarding data conflicts for this traditional abductive inference process. Rather than gathering a collection of data, determining what hypotheses would explain the data, and comparing the plausibility for different combinations of hypotheses in order to come up with a best explanation, the study participants appeared to be following a different process. The main difference between the theoretical pattern of abductive inference and the empirical evidence was that the study participants were not dealing with elemental observations and hypotheses. They were dealing with a "second order" set of data where interpretive frames already existed in which the report writers assumed particular hypotheses and presented data mainly in support of these hypotheses. The main task of the study participant, therefore, was to improve the veracity of the analytic product by corroborating multiple reports of others who had already performed the task of mapping explanatory hypotheses to a dynamically changing data set.

Given this situation, the "hypothesis space" for the simulated task was better represented by Figure 15 than Figure 14. Rather than the "elemental" hypotheses and data given for the Ariane 501 scenario, the think-aloud protocols gave evidence for the study participants dealing at the "second order" level of using cues from the text, document, and source to evaluate how to resolve data conflicts. The study participants displayed expertise in recognizing the cues that were used in evaluating the information and in relating those cues to possible hypotheses.[3]

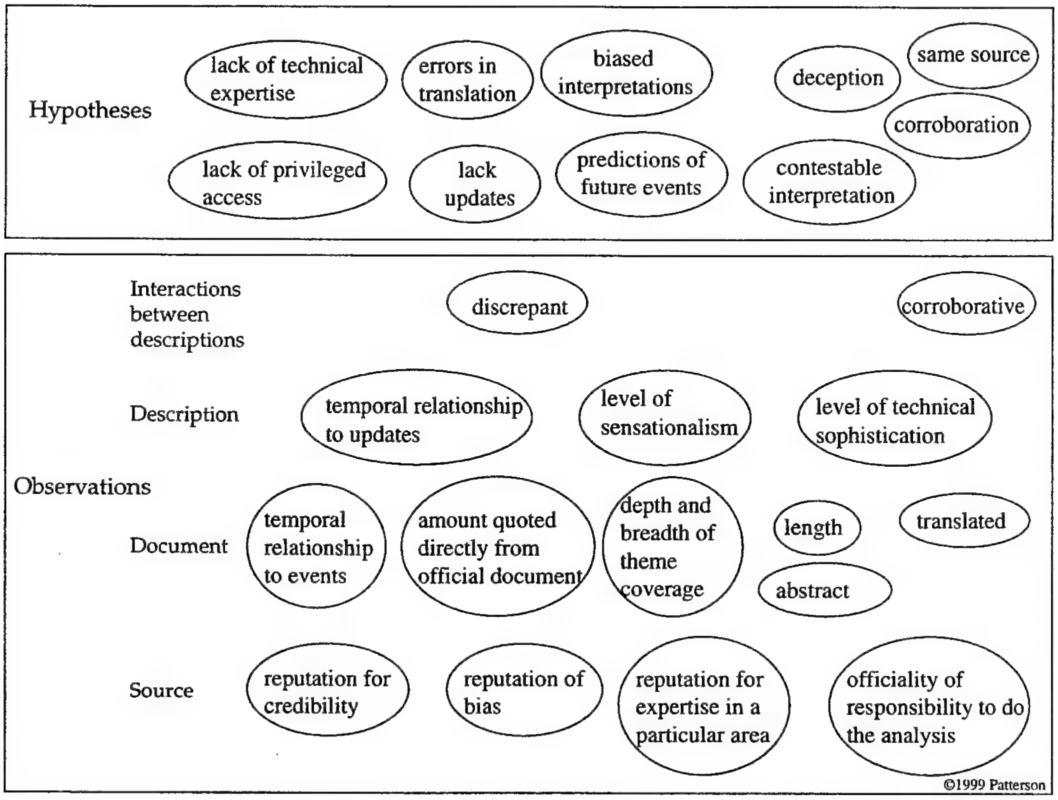

Hypotheses
lack of technical expertise
errors in translation
biased interpretations
deception
same source
corroboration
lack of privileged access
lack updates
predictions of future events
contestable interpretation
Observations
Interactions between descriptions
discrepant
corroborative
Description
temporal relationship to updates
level of sensationalism
level of technical sophistication
Document
temporal relationship to events
amount quoted directly from official document
depth and breadth of theme coverage
length
translated
abstract
Source
reputation for credibility
reputation of bias
reputation for expertise in a particular area
officiality of responsibility to do the analysis
©1999 Patterson

Figure 15. "Second Order" Hypothesis Space

Using the abductive inference framework as a conceptual guide, processes that resulted in inaccurate statements in the verbal briefings were examined to better understand the cognitive challenges and potential vulnerabilities. By tracing why the inaccurate statements were made with the process tracing methodology, three sources of inaccurate statements were identified that provide insight into the cognitive demands of inferential analysis under data overload: 1) relying upon assumptions that would normally be correct, but did not apply in this situation, 2) repeating information that was inaccurate in a document that they had read, and 3) relying upon information that was considered accurate at one point in time, but then was later overturned in subsequent updates.

[3] Note that this expertise would probably not be available to surrogate participants such as undergraduate students.

3.3.4.1 Relying on assumptions that did not apply.

One source of inaccurate statements during the analysis process was the study participants relying on default assumptions that did not apply in this scenario. There were several inaccurate statements made during the verbal briefings that did not come from any of the documents that were opened. For the majority of these cases, the participants appeared to be relying on assumptions to fill in gaps in the story that did not apply in this case. For example, during the verbal briefing, one participant stated that the monetary loss of the Cluster satellite payload could be recovered by insurance. Although payloads are often insured, in this case the Cluster satellites were not.

Relying on assumptions is clearly a heuristic that can be applied under time pressure as a coping strategy. Although relying on assumptions led to inaccurate statements in some instances, in other cases it did not. For example, in one case, participant 2 used the assumption that the Ariane 5 rocket would eventually replace the Ariane 4 as the standard launch vehicle in his estimation of the impacts of the failure. In addition to filling in gaps in knowledge, default assumptions also proved valuable in knowing what information to seek during the analysis process. For example, participant 4 stated that he assumed that there were payloads on the flight and then looked explicitly to see if there were.

3.3.4.2 Incorporating information that was inaccurate.

The second main source of inaccurate statements was inaccurate descriptions in documents in the database. Intelligence analysts clearly view the elimination of inaccuracies by finding converging evidence across independent sources as a major component of the value of an analytic product. The participants described and employed a variety of strategies for tracking and resolving discrepant descriptions in order to reduce their vulnerability to incorporating inaccurate information. Partly because this cognitively difficult process of corroborating information and resolving conflicting information was unsupported by the tools that they were provided, nearly every participant experienced some breakdowns in this process. Breakdowns included failing to corroborate information, missing conflicts in documents that were opened, forgetting how many corroborating and conflicting descriptions had been read from independent sources, forgetting the information sources, and treating descriptions that stemmed from the same source as corroborating (cf., Schum, 1994, evidence interactions in inferential analysis).

To illustrate some of the difficulties in the process of eliminating inaccuracies, consider the example of determining the cause for why the rocket swiveled abnormally. Interestingly, participants 6 and 7 both read the same two documents that contained discrepant descriptions but ended up with different outcomes in their verbal briefings (Figures 16 and 17).

Participant 6 based his analysis of why the rocket swiveled mainly on report 858, which described the cause as a reset of the inertial reference frame following a numeric overflow (Figure 16). As he read 858, he was verbalizing why the rocket swiveled based on what he was reading. Later, he read 1385, which had a contradictory description of why the rocket swiveled. At that point in time, however, it was the last document that he looked at, and he was focused on a different issue – why testing did not reveal the software error. He gave no evidence that he recognized the conflict. In addition, when asked how he knew when to stop the analytic process, he explained: "It doesn't look like anybody will have any different opinions. From looking at the other titles, it looks like I won't come up with anything new."

Therefore, not only did this participant not explicitly conduct the step on this item of corroborating the information through an independent source; he also did not recognize a conflict in what he read. This indicates that recognizing conflicts is a non-trivial task. Direct attention must be given to interpreting that item of information, remembering what had been read in other articles, and recognizing that the descriptions are incompatible. In the electronic environment, this task is particularly challenging because only one report can be viewed at a time because of space limitations on the computer screen. Furthermore, the participant was unaware of conflicts in data that he had read, and as well had no way to tell if there were conflicting descriptions in data that he had not looked at, or even in the reports that were not returned from his query but available in the database.

Participant 6 Briefing: "that guidance system, the length of time that it operated, actually interfered with the inertial guidance system which took over after the launch and it confused...they confused each other and decided that they have to reset but by that time the rocket wasn't vertical anymore"

Article Date/Content	*Participant's Response*
July 5, 1996 (Report 858): *Ariane 5 lifts off much faster... information... exhausted the temporary memory (buffer) capacity...both systems simultaneously declared themselves to be in an irredemiable error situation and commenced a reset procedure...when the system was reset, the vehicle's position at that time...was adopted as the reference base*	"It's the same system as used on the Ariane 4, but the Ariane 5 takes off faster, much faster, than the Ariane 4. The two inertial guidance systems confused each other. They tried to reset at 37 seconds. It wasn't vertical anymore. It just totally lost its mind...so it couldn't figure out its direction."
September 16, 1996 (Report 1385): *the active inertial reference system transmitted essentially diagnostic information to the launcher's main computer, where it was interpreted as flight data and used for flight control calculations*	(talks about a different issue - how it could have been avoided through testing)

Figure 16. Participant 6's Process Trace on Why the Rocket Swiveled

In contrast, participant 7 described the cause of the abnormal rocket swivel as diagnostic information interpreted as command data (Figure 17). This explanation was incompatible because participant 7's description said that there was no command data at all because the guidance platforms had shut down whereas participant 6's description said that there was command data, just that it was incorrect because the guidance platforms had been reset mid-flight.

Participant 7 recognized the conflict in the descriptions in documents 858 and 1440 and resolved it based on a judgment of source quality. He decided to base his analysis on the description in 1440 because it was later and therefore more likely to have all the information, not translated, and from a more authoritative source. Note, however, that even though this was the accurate judgment to make, he did not notice that a previously opened article corroborated the hypothesis that he selected, which would have made the judgment easier. This would have been particularly helpful in this case because, as he pointed out: "[The inaccurate description] sounds good." The description that was inaccurate was written in a way that sounded as if the reporter had sufficient technical expertise to understand the cause in detail. If he had only read article 858 and not found the conflicting descriptions, it is likely that he would have believed the inaccurate description.

Participant 7 Briefing: "numerical values beyond the programmed limits of the flight computer…the platforms initiated a diagnostic "reset" mode that fed incorrect values to the flight computer"

Article Date/Content	*Participant's Response*
September 16, 1996 (Report 1385): *the active inertial reference system transmitted essentially diagnostic information to the launcher's main computer, where it was interpreted as flight data and used for flight control calculations*	nothing
July 29, 1996 (Report 1440): *as a result of the double failure, the active IRS only transmitted diagnostic information to the booster's on-board computer, which was interpreted as flight data and used for flight control calculations*	"We know there was a problem because the guidance platforms shut down. After they shut down, the inertial reference system sent diagnostic information so they're designed to shut down when something goes wrong. Assuming the other system has taken over, it's sending diagnostic information so that the people on the ground can figure out what went wrong with it. Having them both shut down, the guidance computer is interpreting the diagnostic information as where it's at and instead of getting numbers, it's getting other things…"
July 5, 1996 (Report 858): *Ariane 5 lifts off much faster… information… exhausted the temporary memory (buffer) capacity…both systems simultaneously declared themselves to be in an irredemiable error situation and commenced a reset procedure…when the system was reset, the vehicle's position at that time…was adopted as the reference base*	"...In this article, it says when it shut down, it started a reset procedure. In the other article, it says diagnostic information. This article and the other one…are incompatible, inconsistent with each other…Of course messages that can't both be right happen all the time. I'm finding it hard to believe that the vehicle is going to fly without any inertial inputs whatsoever …let's look at the source…FBIS report. Translated text…the other one was later also…it sounds good. If I had to guess, I would go with the other one.

Figure 17. Participant 7's Process Trace on Why the Rocket Swiveled

It was a surprising finding that most of the study participants did not consistently employ strategies to reduce inaccuracies in their analytic products during the simulated task. For example, Guerlain *et al.* (1999) have described that expert blood bankers in antibody identification collect independent, converging evidence to both confirm the presence of hypothesized antibodies and to rule out all other potential antibodies. When asked, the study participants described and, in some cases, demonstrated strategies to protect against the vulnerability of incorporating inaccurate information in their analytic products. On the whole, however, the study participants did not use or only used greatly reduced versions of these strategies during the simulated task, and similarly described that under high workload conditions they tended to do this in the workplace as well. One likely explanation is that the strategies were highly resource-intensive, such as printing out and iteratively using highlighter pens on specific themes to check that information was corroborated from multiple, independent sources. In addition, these strategies were generally not easy to perform within the electronic environment. These observations point to design concepts that would allow the easy manipulation, viewing, and tagging of small text bundles, as well as aids for identifying, tracking, and revising judgments about relationships between data.

3.3.4.3 Relying on outdated information.

The third source of inaccurate statements was outdated information that once had been considered correct but then later had been overturned when new information became available. This type of "inaccurate" information was much more difficult to detect and resolve than misunderstandings by report writers. There were descriptions that were considered accurate at one point in time but that greatly differed from updated descriptions at later points in time. Because the "findings" or data set on which to base an analysis came in over time, there was always the possibility of missing information that was released after the report that was being read that could overturn or render previous information "stale." This occurred both for descriptions of past events where the information about the event came in over time as well as for predictions about future events that changed as new information became available on which to base the predictions. When these updates occurred on themes that were not central enough to be included in report titles or newsworthy enough to generate a flurry of reports, it was very difficult to know if updates had occurred or where to look for them.

To illustrate how easy it is to fall prey to relying on outdated information, consider the process that study participant 6 employed (Figure 18) to come to the conclusion in his verbal briefing that the Cluster satellite program had been discontinued as a result of the Ariane 501 accident: "The immediate impact were that the solar wind experiment was destroyed. They couldn't afford to build any more satellites so they couldn't pursue that anymore." From a global perspective, this is an inaccurate statement given that later updates overturned this initial assessment of the impacts and the Cluster satellite program was later fully reinstated.

Essentially, participant 6 did not open any documents that contained updates on the impact to the Cluster satellite program. The participant opened seven documents during the analysis. Only two of the documents contained descriptions that predicted what the impact to the Cluster satellite program as a result of the Ariane 501 failure would be. In the first description, a scientist working on the project directly stated that the project would be discontinued. While reading this report, the participant verbalized that the scientific mission was dead and that the experiment was destroyed. The second description was more vague about the impact and does not directly make any predictions but could be viewed as weakly converging evidence that the Cluster satellite program would be discontinued. It is no surprise given this process that the participant included in the verbal briefing a description similar to the one from the June 5, 1996 article that the experiment was destroyed and that the program would no longer be pursued. In this case, the participant employed the strategy of corroborating information from two independent, authoritative sources (which would have eliminated the first two sources of inaccuracies), incorporated it into the analysis, and yet missed later updates that rendered that information inaccurate.

Article Date/Content	*Participant's Response*
June 5, 1996: *one of the scientists involved in the project said that it was now finished..."There is neither time nor the money to build four more...the mission is dead, dead, dead."...scientific missions tend to be one-offs and therefore irreplaceable..."All our work just gone in seconds."*	"It wasn't insured...Immediate impact is it was carrying four solar wind experiments and the scientists say that's it, that's all it says, satellites like that are very expensive. The mission is dead, dead, dead...just lost a few satellites. The only immediate impact was that it...and destroyed the experiment."
July 5, 1996: *Why were the cluster satellites, one of the most original, interesting, and costly missions in the space programs, carried on a test flight?...1.8 trillion life for the cluster satellites...down the drain*	nothing

Figure 18. Participant 6's Process Trace on the Impact to the Satellite Program

As a result of basing an analysis on "stale" information that had been turned over by later updates, study participants made several inaccurate statements at varying levels of importance. The vulnerability to missing critical information is particularly troubling because it is so difficult for practitioners to determine when they have missed critical information. It is the *absence* of information, either from not sampling the information or having attention directed on a different theme while reading a document, that creates the vulnerability.

3.3.5 Summary of Observed Behavior and Design Implications

By observing expert intelligence analysts on a relatively complex, face valid task using a baseline set of querying and browsing tools similar to what is available to them in their workplaces, we were able to greatly increase our understanding of the challenges of intelligence analysis. Under the extreme conditions of a short timeframe of several hours in a new topic area with a database and question unfamiliar to the analysts, we observed behaviors across most or all of the study participants that pointed to design recommendations (Table 5).

Table 5. Summary of Observed Behavior and Suggested Recommendations

Observed Behavior	Suggested Recommendations
All participants did not characterize the database available to them	Information visualizations that allow interactive, real-time exploration of the characteristics of subsets of data
All participants narrowed in on a small set of documents	Reminding functions to explore other portions of the database; Visualizations that allow natural browsing of larger sets of documents
All participants appeared to read the first documents they opened more carefully than later documents	Better algorithms to identify high quality documents; Support for identifying data conflicts and updates; Support for identifying new information in a document
All participants did not conduct new searches	Machine suggestions for additional queries to perform; Improved usability of query formulation functions; Visualizations to allow comparisons of search results; Visualization of browsed documents against the available set
Participants who made inaccurate statements did not read high profit documents; All participants missed some high profit documents	Machine suggestions for candidates of high profit documents; Training on characteristics of high profit documents; Support for locating similar documents to a tagged set; Design of interface to encourage narrowing by document attributes not keywords
Some participants missed important events	Information visualizations that enable event recognition; Machine processing to highlight possible events based on heuristics
Some participants missed data conflicts	Support for identifying and tracking data conflicts; Training on identifying and resolving data conflicts
Some participants could not remember where data came from; Time-intensive strategies to track source information	Support for identifying source documents; Identification of duplicate information from the same source
Some participants missed important updates	Support for locating and tracking updates on a theme
Wide variation in confidence estimates about accuracy of verbal briefings	Make it easier to identify when events, conflicts, and updates might have been missed; Reminder functions for data conflicts; Visualization that allows holes in analysis to be made visible

First, several of the study participants expressed uneasiness because they were unaware what was potentially available in the database provided to them. In addition, an expert analyst provided insight that it was interesting that none of the study participants explicitly attempted to characterize the database at any point during the analysis by performing multiple queries to see what was returned. The desire to evaluate the quality and type of information that is returned by a query against what is potentially available might explain why all analysts create personal databases on topic areas for which they are responsible. When analysts are then asked questions about a new topic area, they lose this ability to calibrate expectations about what is returned in comparison with what is potentially available. These observations point to several ideas for design recommendations. Specifically, information visualizations could be created that would allow interactive, real-time exploration of the characteristics of subsets of data. Although there are several software packages that exist that attempt to do this, the only feedback about the characteristics of the dataset returned in the tools provided to the study participants was the number of returned hits.

Second, it was observed that all of the study participants narrowed in on a small portion of the dataset and performed all of their further searches for information from moving within that space. This observation leads to possible recommendations to encourage analysts to explore other portions of the database, either by explicit machine recommendations or through interface designs that naturally suggest how much of a set of potential data has been explored. In addition, visualizations that allow easier browsing of larger sets of documents could make the set that the analysts narrow to be larger and thus inherently more of the database is covered in the sampling by dates and titles.

Third, it appeared that there was an interaction between the order the documents were selected from the browser window and the time and effort spent reading the document. The first or second document that the analyst selected for reading in detail seemed to frame how the rest of the information was later interpreted. This observation indicates the importance of quickly providing high quality documents to an analyst. In addition, when later documents are quickly browsed, data conflicts, updates, and new information could somehow be highlighted to reduce the chances it would be missed.

Fourth, although several study participants verbalized that they should conduct a new search, (s)he did not and appeared reluctant to leave the working area that the home query window represented. The participants developed a familiarity with the titles and dates of the documents returned by the query, the participant had often sorted the documents by date, the windows had been resized and placed in a dedicated place on the screen, and some of the documents had been marked for various reasons. This observation leads to recommendations for passing highlights and "trails" of what had been opened to new query returns.

In addition, query formulation was relatively difficult to manipulate in the interface, and so performing "what if" changes to a query formulation would require forming separate queries for each formulation and then comparing the number of hits returned.

Fifth, the study participants who located "high profit" documents made fewer inaccurate statements in their verbal briefings than those who found none. If, in fact, the explanation for the difference between the two groups is the amount of time and the number of documents, then this indicates that one of the ways, given a baseline electronic toolset of keyword querying and browsing by dates and titles, to find the high profit documents in the database might be to cast a wider net by sampling more, either by performing more queries or by opening up more documents. Support tools such as "agents" that remind or critique analysts to be broader in their sampling strategies might be helpful. However, given the increasing organizational pressures to do analyses more efficiently, these types of support tools might be ineffective because analysts might not have the resources to do so. A potentially more viable design intervention to reduce the vulnerability to missing high profit documents would be to use machine intelligence as a "recommender" system to suggest likely candidates for high profit documents. For example, for this scenario, a high profit could be characterized as: 1) a relatively long document that was released several months after the original event (and certainly after the Inquiry Board Report was officially released from the European Space Agency), 2) from a credible source on rocket launcher and satellite technologies such as Aviation Week and Space Technology, 3) not an abstract, 4) not reporting information from another news agency (i.e., not "secondhand"), 5) not translated from another language, and 6) a report that had been opened several times by others.

Sixth, some participants missed important events in searching for information, such as the launch of the next rocket in the series, Ariane 502. It was observed that, in the documents returned by the study participants' queries, there were clusters of reports around the time of the 501 rocket launch failure, when the Inquiry Board Report was released, and the next launch in the series, 502. This observation led to the idea that disrupting events could be visually emergent from a display, becoming an implicit cue where an analyst should look for informative data.

Seventh, breakdowns were observed in the process of resolving discrepancies in the data, such as failing to identify discrepancies in information that was read and double-counting information from the same source. These observations led to the concept of aids for identifying, selecting, manipulating and tracking judgments about conflicts and corroborations in data.

Eighth, many study participants were observed to devote considerable time and effort to methodically tracking what document information came from (e.g., copying source information in a word processing program into footnotes associated with text selected from a particular document). In some cases, study participants were observed to state that they forgot where information came from or that they were uncertain if information was new or a repeat from reading the same document again.

Ninth, study participants were observed to make inaccurate statements because they missed updates that overturned information that was once considered accurate. In addition, it was observed that many of the study participants had difficulties in identifying discrepancies in predictions about when events would occur from text descriptions such as "a few months from now" from one report at one time and "delayed for several months" from another report at a different time. This observation led to the notion of visualizing this information on two parallel timelines, connecting the document date on one timeline with the predicted event date on another to facilitate recognizing patterns such as conflicting predictions and slips in predicted times. Aids that would remind users to search for updates and suggest possible areas to look for updates based on similarity matches to text descriptions and other attributes could potentially be very useful.

Finally, we were surprised by the wide variation in answers about accuracy as estimated by the study participants immediately following their verbal briefings. It appears that it is extremely difficult to determine a sound basis for a confidence estimate given that there could always be information that was missed that would greatly alter or overturn the analysis. Analysts clearly need support in identifying potential "holes" in the analysis process, due to both missing information and leaving issues unresolved. Visualizations that represent the state of the analytic process might help improve analysts' ability to calibrate their assessment of their accuracy, including displays that show what information has been sampled and assembled together, as well as information that has been "tagged" or "bookmarked" as a reminder to return to resolve open questions.

As we had learned previously in interviews, the observed behavior during the study indicated that the baseline computer support tools left most of the challenging tasks in conducting analyses under data overload conditions unsupported or only weakly supported. We believe that the relationship between the challenges in inferential analysis based on sampling uncertain and conflicting data and the support provided by the baseline electronic environment is likely the primary explanation for these patterns of observed behavior across study participants. The observed behavior left the study participants open to making incomplete and inaccurate statements in their verbal briefings. These observations point to new directions for computerized support for these processes.

3.4 Developing Evaluation Criteria

The findings from the study provide insight into what cognitive demands in supervisory control under data overload are most prominent in intelligence analysis. By conducting the study, we were able to more directly target designs that would be useful to the analysts in that they would reduce vulnerabilities to generating inaccurate or incomplete analytic products. If we had not conducted the study, we might have designed systems that might have appeared innovative in a demonstration and could be useful, but that would likely have incorporated features that would be infrequently used, thereby creating unnecessary complexity in the interface and expense in the design process.

In addition to generating new design concepts that we pursued, the study also allowed us to translate the identified vulnerabilities into specific criteria that successful responses to the data overload problem in intelligence analysis need to satisfy. Therefore, this step also had the benefit of creating criteria that could be used to objectively evaluate the usefulness of any design concept designed to combat data overload in intelligence analysis.

1. *Recognition of Unexpected Information.* Bring analysts' attention to highly informative or definitive data and relationships between data, even when the practitioners do not know to look for that data explicitly. Informative data includes "high profit" documents, data that indicates an escalation of activities or a disrupting event, and data that deviates from expectations. A particularly difficult criterion to meet that should be designed into evaluation scenarios is to help analysts recognize updates that overturn previous information.

2. *Management of Uncertainty.* Aid analysts in managing data uncertainty. In particular, solutions should help analysts identify, track, and revise judgments about data conflicts and aid in the search for updates on thematic elements.

3. *Broadening.* Help analysts to avoid prematurely closing the analysis process. Solutions should broaden the search for or recognition of pertinent information, break fixations on single hypotheses, and/or widen the hypothesis set that is considered to explain the available data.

These evaluation criteria are interesting, in part, because they are so difficult to address. We realized quickly that these criteria are not amenable to simple, straightforward adjustments or feature additions to current tools. Meeting these design criteria will require fundamentally innovative and novel design concepts.[4]

At this point in the project, we took stock of what our methodology had provided us as a research/design team. We believed that already at the completion of the study, we were able to see progress that reinforced our belief in complementarity between research and design and learning about understanding, usefulness, and usability in parallel because:

- it contributed to our general understanding of data overload, as evidenced by helping us in other settings such as NASA Space Station mission control,
- it revealed the world of the analyst effectively and grounded general concepts to the particulars of the situation the professional analyst faces,
- we were able to identify characteristics of intelligence analysis that were similar and unique to other settings in which we had more experience,
- generative design sessions had a step upward in productivity and we were able to eliminate directions and features to pursue and come to better consensus within the team as to what concepts to emphasize,
- we were better able to critique and offer suggestions to improve ongoing projects aimed at solving the data overload problem,
- it generated new practice-centered criteria for evaluating proposed solutions to data overload,
- it is serving as a basis for interaction and as a stimulus to a more constructive dialogue across analysts, developers and others for useful design directions to pursue,[5]
- many in the analyst community could take home lessons for their own role or work. For example, a spin-off project at the agency is underway where the Ariane 501 scenario and database will be used as a training vehicle for new analysts while they are waiting for their clearance.

[4] We would like to note that these criteria, although so easily recognized in hindsight that they have been challenged as obvious, are different than criteria that were described to us previously. For example, prior to conducting the study, criteria for addressing the data overload problem were offered at various points to be 1) to have an analyst be able to read it all, 2) to find the relevant information that is needed to perform an analysis, 3) to visualize the landscape of the information space, 4) to have the machine tell an analyst when an important message has been received, 5) to see an overview of events in an area that have not been monitored for some time, and 6) to have the machine summarize the important points in each message.

[5] It was fascinating to watch a developer and an analyst interact around the kinds of concrete issues that the study captured after one of our presentations.

PART IV. EXPLORING DESIGN SEEDS

Many Research and Development (R&D) projects would transition to a development team at this stage. In a traditional situation, the development team would brainstorm different design approaches to take based on these research results, select a particular direction, and begin iterative cycles of developing, evaluating, and implementing a design concept. The types of questions that would be asked during these stages might include:

- Is this software better than the alternatives?
- Does the software work as intended?
- Is the software easy to learn and to use?
- Is the software cost-effective? Is there a market for this software?
- How should the software be sold?
- Will the software require new infrastructure or can it be integrated into existing hardware and software infrastructure?

In our conceptual framework of levels of design concepts in Figure 5, most of these questions would be at the "Usability" level of design. Instead of being a diamond shape where explorations of the usefulness of design concepts are emphasized, the shape would be like an hourglass. Research is performed that clearly increases the understanding of the challenging tasks in a domain and a workable design concept is made usable and perhaps even implemented, but there is little to no evaluation and iteration on making a design concept usefully address challenges to real-world practitioners. The risk is that a workable, usable design concept will be implemented only to be ignored by users in the field who feel that the aid provides insufficient support to meet the requirements of their roles and responsibilities.

Our stance is that the translation from a solid understanding of the challenges faced by real-world practitioners to the creation of a design concept that is useful in addressing those challenges is an effortful, difficult, and under-appreciated step. Even following extensive research, it is likely that the first attempts will be inadequate and require large fundamental shifts in the conception of what would be useful to design.

There are several reasons for initial attempts at design concepts based on high-quality research to fail the "usefulness" criterion. One of the primary reasons is the difficulty of moving from an understanding of the current challenges in a domain to predicting how a design will support practitioners in an envisioned situation that will be different from the current situation. The introduction of new support tools will fundamentally alter the tasks of the domain. In addition, domain challenges are a moving target from organizational, design, and training shifts constantly underway. Note that real-world practitioners are also not able to adequately address these challenges: their expertise is based on their current and past situations, not about predicting how other experts will act in the future under different conditions. As one expert intelligence analyst wrote us following a presentation of design concepts: "If there is any one single thing I would caution you about relating to the wonderful work you are pursuing, it is not to listen to only ONE analyst...[a software package we use] has been down that path and we are still living with the mistakes of this nature from the past. Oiling the squeaky wheel may relieve the squeaking but it MAY NOT get to a basic underlying problem..."

4.1 Factors Driving Commitment to a Single Design Direction

Following the shift from a research phase emphasized at better understanding the challenges of a domain to a design phase, there is often a rush to commitment to a single design direction. The outcome of this commitment is often an inability to reorient design concepts at the usefulness level based on feedback during evaluations. One of the methodological contributions of this project is an illustration of Cognitive Systems Engineering (CSE) techniques aimed at reducing the risks of devoting expensive resources to developing and implementing a design concept that will be ultimately thrown away or under-utilized by real-world practitioners.

There are many factors that drive early commitment to a single design concept (Table 6). Organizational resource allocation procedures often inhibit development teams radically shifting from one design concept to another based on user feedback. Similarly, the structure of many R&D organizations is designed around the idea that research results from cognitive task analyses are fed to development teams, at which point the research ends and the development efforts begin. Even in environments where the exploration of the usefulness of design concepts is valued, deadline and workload pressures can drive out the ability to explore broad regions of possible design concepts. In addition, in order to maintain organizational investments in a design concept, demonstrations are often used to monitor progress. In preparation for these demonstrations, often choices are arbitrarily made that could in theory be later overturned. However, most of these choices are never later re-examined. Finally, one of the most impressive drivers of early commitment to a single design concept is the emotional and psychological investment that a team creates through the process of designing it. Even when no other pressures are on a team to commit, such as with educational design projects, it is surprising how difficult it is for design teams to redirect in the face of evidence that a concept might not be useful to practitioners in the target setting.

Table 6. Factors That Drive Early Commitment to a Single Design Concept

Organizational	Structure of development team projects Structure of R&D organizations
Deadline pressure	Need to meet deadlines drives out ability to change course
Workload	Effort to maintain multiple concepts in parallel
Economic	Cost to maintain investment in multiple concepts
Demonstrations	Developing demonstrations forces choices that are not later re-examined
Psychological	Emotional investment in a concept Feeling of progress toward a fielded product

4.2 Design Seeds

A common problem in R&D organizations is the transfer of a design concept from a research to a development project. Often, a (partially) working prototype is used to "embody" the requirements for the design of a working system. This working prototype is often compared in a "head to head" evaluation with the currently available system to justify implementation. Nevertheless, when an actual product is fielded, many other constraints need to be taken into account, such as interactions with other software, existing infrastructure, implementation cost for particular features, and reliability and speed criteria. Because the design concept is embedded within an integrated whole, it is difficult to determine how to trade off with the other interacting constraints while still maintaining an effective system.

A similar concern in transferring research to development is how to take advantage of research innovations without beginning a completely new development project. A common question is what portions of a design concept would be useful for existing development efforts to incorporate from research efforts on shorter timeframes than would be required to create the entire prototype.

To address these problems, we have developed the notion of modular "design seeds." A design seed embodies a particular concept of what might be useful to support human performance that could be instantiated multiple ways in different design efforts and does not depend on other design seeds to work. The design seed represents a target for a design team to shoot for as well as something that can be individually explored for how useful it is for users in research studies. Design seeds do not necessarily map one-to-one onto functions or interface elements – they represent strategies for aiding users for a particular leverage point that we believe are likely to be effective based on research findings.

Note that, to take advantage of and invest in research bases about useful design seeds in cognitive systems engineering, there is another important requirement for design seeds. Descriptions of 1) the design concept, 2) the vulnerability the concept is trying to address or the strategy the concept is intended to support, 3) criteria for how weakly or strongly to commit to results of machine processing, and 4) expectations on impacts to human performance should be generalizable to other domains. Without this abstraction and dissemination of design seeds, each development project will be forced to revert to expensive trial and error because they will be unable to take advantage of lessons learned by others.

4.3 Jumpstarting Design With Research Base

When we shifted our focus to exploring design seeds, an explicit step in our methodology was to cull ideas of what might be useful design strategies from available resources in order to jumpstart our ideation of design seeds. Although we discussed possible design concepts right from the first meeting of the project, it was at this point that we explicitly surveyed, expanded, and synthesized what we could use as a foundation. We drew on research bases[6] and our own past experience in designing visualizations in other domains, particularly in space shuttle mission control, nuclear power plant control rooms, and critical care medicine.

Based on this research base on how to design useful systems, we focused our efforts such that:

1. we identified two observable, interacting frames of reference (time in "report space" and time in "theme space") as the base structure for the main display,
2. we coordinated the elements of the workspace to function as a unit in ways that could be viewed from multiple perspectives,
3. we made the interface observable and directable,
4. we designed a series of "longshots" from which to view large-grained patterns in the spaces and reduce the navigation burdens,
5. we considered how to use the machine to critique the activities of the human partner, and
6. we circumscribed the active machine intelligence such that the system is not dependent on the machine processing always being correct.

[6] Although we feel that we drew on research bases about design concepts, we are unable to reference archival publications that explicitly describe this research base. Traditional publication outlets generally do not accept descriptions of explorations and descriptions of useful design concepts as worthy of publication in their own right. Traditional design guidelines often do not emphasize the level of how to design systems to be useful. Therefore, we feel that much of this information is learned via apprenticeship or in informal discussions with members of research and design communities.

4.4 Animocks: Animated Fly-through Mockups

As with all of our research and design activities, we feel that there are multiple conceptual levels embedded within objects (Figure 5). In order to evaluate the usefulness of a design seed, it needs to be instantiated in some form. Because we wish to reduce commitment as much as possible to a design direction so that we can be more responsive in the face of evidence that a concept is not useful, we recommend instantiating design seeds in "throw-away" formats that are based in scenarios and can be quickly prototyped without requiring all of the interaction and interface details to be worked out. When instantiating design seeds, we are definitely engaged in a design endeavor that explicitly forces consideration of some of the most important design trade-offs without committing to a particular design. The instantiation of a design seed focuses design discussions in a new area of a design space rather than encouraging incremental advances on current artifacts.

We instantiated our design seeds as animated fly-through mockups in the context of variations on the Ariane 501 scenario, or "Animocks." With Animocks, commitment to a particular hardware infrastructure, visualization instantiation, or combination of design seeds is lessened and there is greater flexibility to incorporate feedback about the usefulness of a design concept in addressing data overload. In addition, the design seeds are designed to be conceptually modular and based in challenging scenarios, which enables generalization of concepts across design projects and domains. With this strategy, we can better address one of the primary challenges of a research and development program, which is to develop research bases that can be translated into fieldable systems in multiple settings to prevent continuously engaging in individual, one-off design endeavors with no learning about how to improve systems over time.

We now describe in detail the design seeds that we generated to address challenges in conducting inferential analysis under data overload.

4.5 Design Seed 1: Exploratory Searching

One of the main vulnerabilities during the simulated analysis task was missing critical information without being aware of it by prematurely closing the search process. All of the study participants quickly narrowed in on a set of documents on which they based their analyses. Although several mentioned that they were concerned that they did not know how their sample related to what was potentially available, none of them conducted further searches or explicitly characterized the database through exploratory searches. Therefore, we have developed a design seed that instantiates support for exploratory searching in order to allow analysts to have a better sense of how their samples relate to what is potentially available to them in a database.

Although there are potentially many ways to implement this design seed in an actual system, several techniques that promote good usability are likely to be used in some way. First, the system must somehow allow users to gain a sense of how the documents they have sampled relate to the space of available information by allowing visual comparison of sampled sets against each other and the entire document space. In an interface with good usability, users will get near-immediate feedback on how changing search parameters and document attributes affects their sampled set. Visualizations based on mechanisms such as procedural search constructions where changes are not made sequentially but can be immediately implemented at any point in the search construction would support this need in a natural fashion.

In summary, this design seed has the characteristics of:

- helping analysts to characterize the information space by allowing real-time "what-if" manipulations on search parameters,
- supporting parallel comparisons of sampled documents against each other and the entire document set, and
- easily manipulable search terms and parameters with immediate feedback on the sampled set.

One example of how this design seed is expected to impact performance includes making it easier for analysts to quickly troubleshoot why no results were returned from a search that added "Aerian" as a search term. The user would see that the addition of Aerian as a search term reduces the returned set to zero, realize that the word is misspelled, and replace it with "Ariane." This scenario is compared with search term constructions that must be completely created and then "submitted" to a database with results returned some time later in the format of "688 hits" or "0 hits." Although in both situations it is likely that there was an erroneous spelling or term in the search construction, in the second case the entire search term is treated as an integrated unit and no clues are provided as to where to change aspects of the query.

Another example of how this design seed might impact performance is by helping an analyst to see how adding the term "1996" to a search construction affects the document set that is sampled. It is possible that adding the search term "1996" would eliminate all documents prior to 1996 and greatly reduce documents after 1996. By experimenting with the interface, the analyst can visually see that there are fewer documents overall compared with the search without the added keyword "1996," and that there are in fact changes to the weightings of the documents in time (Figure 19). Some documents prior to 1996 are still available with the keyword, the documents from 1996 are somewhat increased (from 52% to 64%), and there are fewer documents after 1996 (40% to 22%).

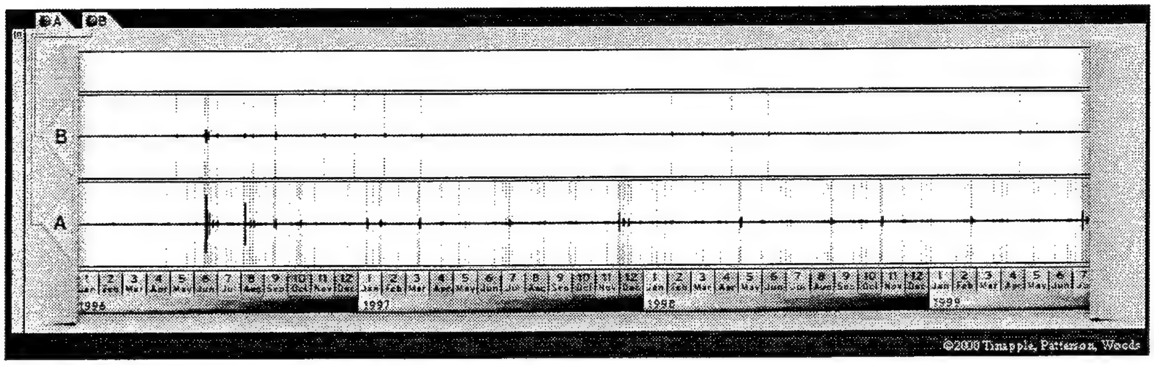


Figure 19. Search Results without (Query A) and with (Query B) "1996" Keyword

4.6 Design Seed 2: Critiquing Search Tactics

When comparing the search strategies of the study participants with tactics described in the information retrieval literature (Bates, 1979), it became clear that the participants were "search novices" in that they used relatively crude search tactics that left them vulnerable to missing important information. Their searching strategies could be improved by:
1) using synonyms of keywords combined with "OR" connections to be more exhaustive in selection of concepts referred to by different words
2) truncating words to capture more forms of a word based on a common root (e.g., los* instead of lost)
3) using orthogonal facets to narrow search results (e.g., crossing organizations with an event such as "ariane" AND "los*")
4) using document attributes rather than additional keywords when facets are no longer orthogonal (e.g., instead of using destr* AND fail* to narrow a search, restrict documents to a certain time window or to documents written in the English language)
5) developing an effective search model and reusing it when beginning a new search (e.g., the countries and technologies normally involved in tasks as well as the sources that tend to be useful in answering different types of questions).

In order to improve the search tactics of intelligence analysts, the machine processing is viewed as a teammate working with the human agent. The design decision of how much the machine needs to initiate vs. follow up human actions depends on several factors. First, does the population that uses the system know about the above search strategies? If not, then perhaps the system should include an educational component about what it is trying to achieve. If so, then perhaps the machine agent could simply remind the human agent to achieve these goals. If there is a diversity of search knowledge among the user set, perhaps the interface could be designed in a way that naturally "affords" these search strategies, such as by providing multiple lines for synonymous ("OR") search terms within columns that represent facets to be "ANDed" against each other.

Second, why do the human agents not employ these strategies in their daily work? If they do not use them because they do not think of them, then perhaps a machine critiquer that pops up dialog boxes with suggestions to improve the search would be useful. If they do not because they do not feel that they have the time to employ them, a system that makes it more time-efficient to conduct better searches might help, such as by providing templates or saving previous searches for re-use. If they are not able to quickly think of synonyms for words, then perhaps synonym dictionaries, either generic or tailored to their areas, might help. If they find it easier to think of words like "lost" than "los*", then perhaps the system could automatically change formulations which the human agent could then override if it is not desired.

Third, if human agents do not employ these strategies because they find it cognitively easier not to do so at the unknown expense of missing critical information, perhaps the machine needs to be more forceful in requiring the human agent to use them. Rather than "weakly" supporting the human by the interface design, reminding, critiquing, or suggesting, as the previously described strategies have, the machine agent might force the user to type in at least one synonym in each facet before conducting a search. Although this type of machine intelligence might prove to be useful, it needs to be designed with great care as it reduces the flexibility of the user to meet unpredictable situations.

In summary, this design seed has the characteristics of:

- providing interface "affordances" that make it easier for users to use effective search tactics
- improving search strategies through machine critiquing, suggesting, and/or reminding
- allowing the reuse of previously effective searches
- helping users to learn more effective information retrieval strategies

An example of how this design seed might work in a scenario includes the following. A user would type into an interface that encouraged synonyms within orthogonal facets the keywords in Table 7, also shown in Figure 20.

Table 7. Search Term Facets for Ariane 501 Entered by a User

Facet 1: Ariane	Facet 2: ESA	Facet 3: Lost	Facet 4: Cluster
Ariane	European Space Agency	lost	launcher
Arianespace	ESA	fail	rocket
Ariane 5		explode	
Ariane V			

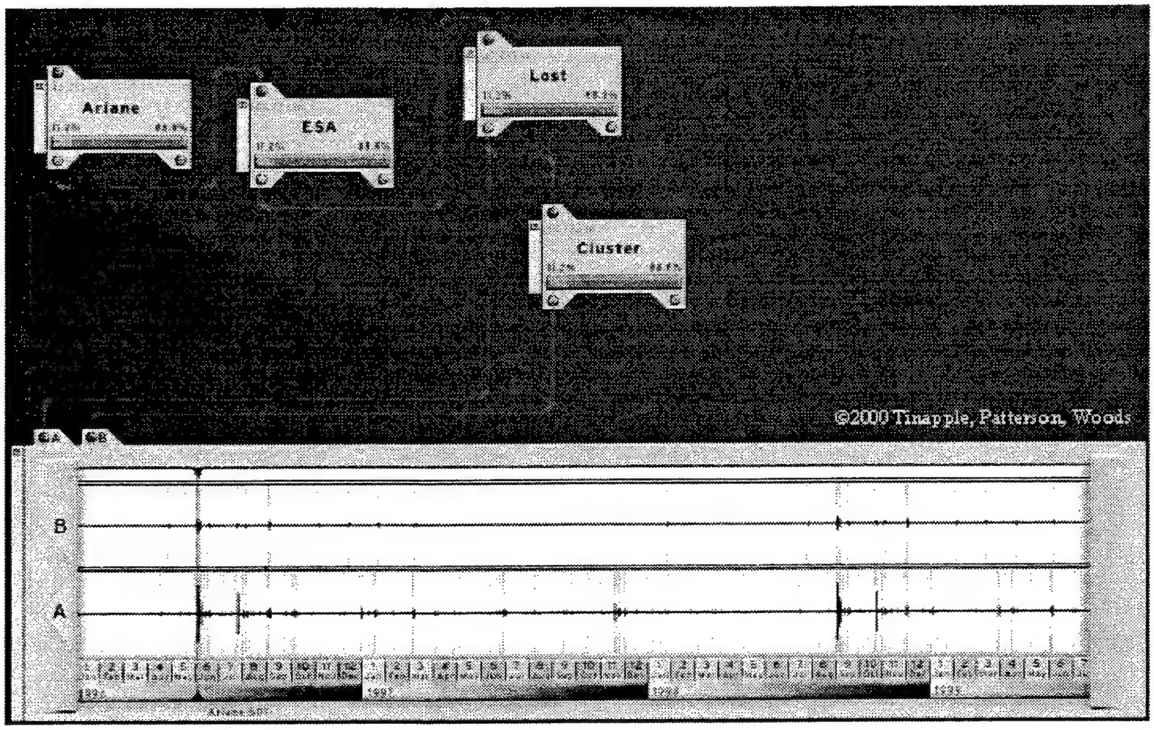


Figure 20. Interface Visualization for Search Term Facets

The software recommends that the user use los*, fail*, and explo* in place of the verbs in facet 3. This can be done either by the machine automatically replacing the terms with the recommended form or by an alert dialog box suggesting that the user accept the changes.

The user judges that he received too many hits with the current search formulation (240 hits). So he adds a fifth facet node and types destroy to "AND" with the rest of the search. The machine critiquer pops up a dialog box that says that destr* is a synonym with an existing facet and so should not be used in this way to search. It lists a set of document attributes that might be used instead to narrow the returned set:

- document date
- a list of sources that the user had previously entered as being "A list," "B list," and "C list"
- number of words
- translated only or NOT translated
- abstract only or NOT abstract
- language.

The user restricts the set of documents to the English language, which then returns 80 hits. At this point, the user begins to select sets by date from areas of the returned set to browse in more detail.

Note that there are multiple ways that the machine intelligence can be used to encourage better search tactics. For example, it is recommended that concepts that are orthogonal to each other (e.g., the rocket launcher and the payload) be crossed with AND combinations while expanding the terms used to search for a concept with OR combinations e.g., "European Space Agency" and "ESA"). The machine processing could be used to encourage this tactic:

- through the design of a visualization which naturally affords typing multiple words as synonyms for the same concept within a facet by having multiple empty rows within the facet,
- by directly suggesting synonyms from a generic dictionary,
- by directly suggesting synonyms from a tailored or personal dictionary based on previous searches,
- by providing advice about effective search tactics prior to or while searching for information, or
- by providing a computerized tutorial or contact information for obtaining training on search tactics that users can access to improve their search effectiveness.

4.7 Design Seed 3: Searching for High Profit Documents

One of the main insights from the study was the impact of basing an analysis on a small number of documents that were judged to be of great importance by the study participants, which we refer to as "high profit" documents. It is interesting to contrast the "design seed" that might have been created relating to high profit documents based on interview data with the design seed that we generated based on the study observations. During interviews prior to the study, analysts described that having high quality information was critical to conducting a high quality analysis. When asked to describe what they meant by high quality information, they were general in their descriptions, for example describing that information came from a "good source" or had a "certain flavor." We believe that an obvious design seed based on this data would allow sorts by source names or filtering by source names, such as by color coding tiers of sources as in Table 8 and the associated visualization in Figure 21.

Table 8. Color-coding Categories for Good, Moderate, and Poor Sources

Best (Green)	Medium (Blue)	Poor (Red)	Unknown (Black)
Aviation Week and Space Technology	Nouvelle Revue Aeronautique	FBIS report	Paris Air & AMP
Air et Cosmos Aviation	Revue Aerospatiale	Guardian Newspapers Limited	Information Access Company
Defense Daily	New York Times	Newspaper Publishing PLC	Foreign Press Survey
Aerospace Daily	Detroit News	Amembassy Paris	Journal of Commerce
Financial Times	Washington Post	AMCONSUL FUKUOKA	Ottawa Citizen
Satellite News	The Times Newspapers Limited		Telegraph Group Limited
			Aero-Espacio

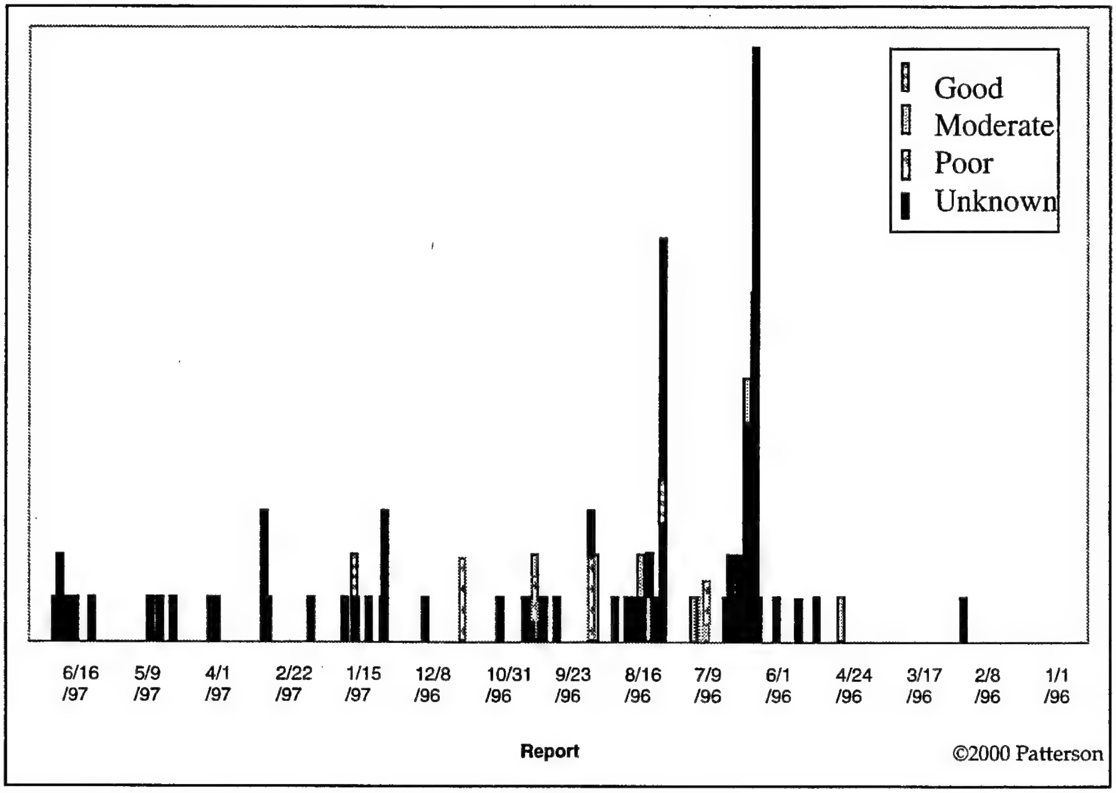


Figure 21. Interface Color-coding Good, Moderate, and Poor Sources

From the verbal think-aloud protocols from the study, we were able to identify more cues in judgments of data quality than from the interviews. During the study, participants took advantage of cues at the level of the document and portions of text in the document in addition to other information about the source (Table 9).

Based on this data, we feel that we can take advantage of more attributes, and particularly combinations of attributes, to suggest fewer and more likely "high profit" candidates to analysts. For example, high profit documents in the Ariane 501 scenario could be characterized as relatively long documents from a small number of sources that were published at least a week after the release of the official Inquiry Board Report (see Table 10 for high profit documents in the study scenario). Similarly, some articles were considered low profit because they were abstracts, translated, secondhand in that they summarized reports from other sources (e.g., FBIS), were from sources that were likely to be biased, or were judged to be sensationalistic.

Table 9. Cues to Data Quality at the Source, Document, and Description Levels

Source	Document	Description
Reputation for credibility	Temporal relationship to events (do not have all the information right away)	Temporal relationship to updates (can be "stale")
Reputation of bias	Amount quoted directly from the official document	Level of sensationalism
Reputation for expertise in a particular area	Distance from the original data: secondhand, translated, summarized	Technical language
If given official responsibility to do an analysis	Length	
	Depth and breadth of theme coverage	

Table 10. High Profit Documents

Date	Title	Source	Number of Words	Abstract?	Translated?	FBIS?
July 19, 1996	Ariane 5 Flight 501 Failure: Report by the Inquiry Board	Inquiry Board	5457	N	N	N
July 24, 1996	Inertial Reference Software Error Blamed for Ariane 5 Failure	Defense Daily	624	N	N	N
July 24, 1996	Software Design Flaw Destroyed Ariane 5; next flight in 1997	Aerospace Daily	629	N	N	N
July 24, 1996	Ariane 5 Rocket Faces More Delay	The Financial Times Limited	411	N	N	N
July 25, 1996	Flying Blind: Inadequate Testing led to the Software Breakdown that Doomed Ariane 5	The Financial Times Limited	984	N	N	N
July 29, 1996	Board Faults Ariane 5 Software	Aviation Week and Space Technology	1501	N	N	N
August 5, 1996	Ariane 5 Explosion Caused by Faulty Software	Satellite News	3762	N	N	N
September 9, 1996	Ariane 5 Report Details Software Design Errors	Aviation Week and Space Technology	2554	N	N	N
September 16, 1996	Ariane 5 Loss Avoidable with Complete Testing	Aviation Week and Space Technology	2834	N	N	N

Because the characteristics of a document that a machine is able to recognize are likely to not match critical documents in every case, we feel that it is important to use this "model" of characteristics of a high profit document in a way that does not rely heavily on the machine processing. For example, a poor implementation of this model would be a computer system that provided what it thought candidates for high profit documents might be without displaying the criteria it applied, allowing the user to change the criteria, allowing the user to see how the set compares against other sets and the entire database, and particularly one that would force the user to open each candidate based on the principle that users would then be less likely to miss critical information.

In contrast, we believe that human-computer architectures that are less strongly committed to a specific model of a "high profit" document might prove useful. For example, a support system that leveraged the model of a high profit document could be implemented different ways depending on how effective the algorithm is at identifying all of the likely candidates for high profit documents without also identifying many other lower profit documents in the process:

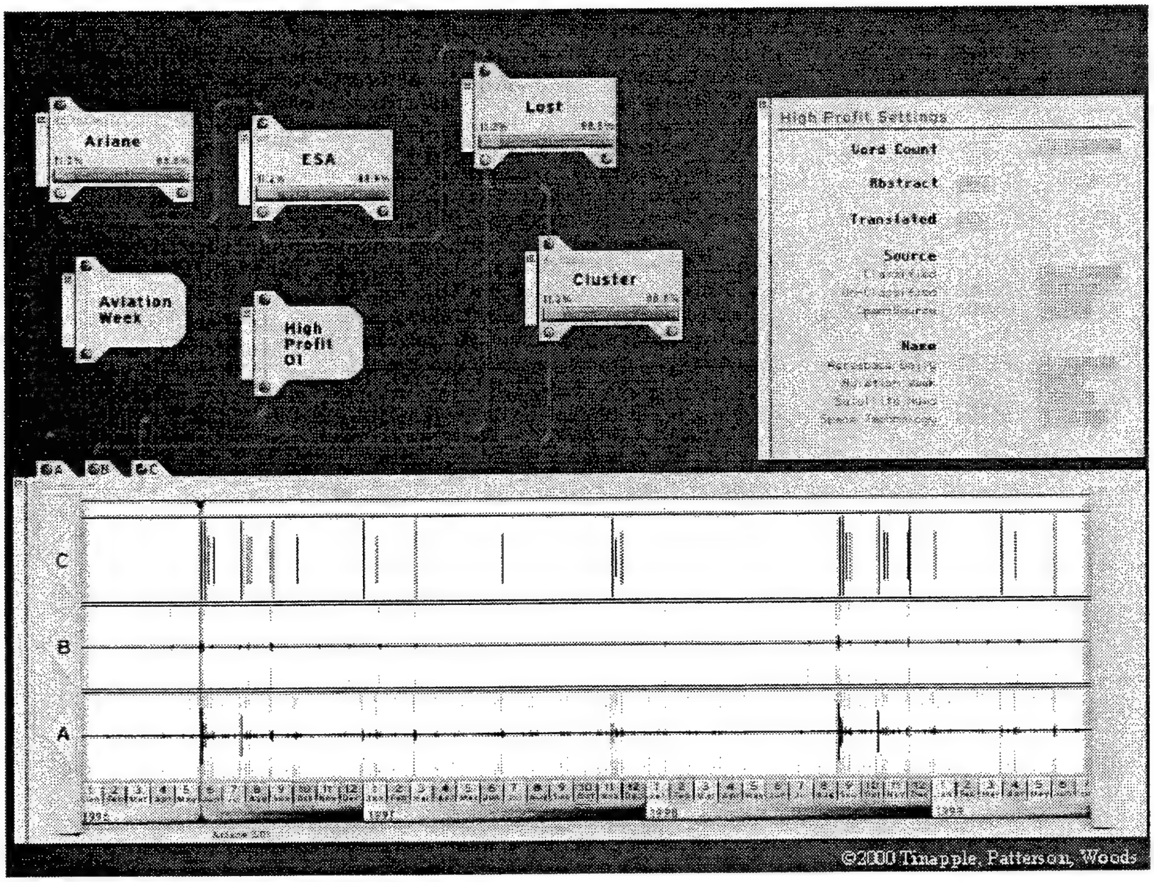
1) the user could mark a document as high profit and the computer could then display and categorize that information in various ways,
2) a computer algorithm could determine similarities in documents that were marked as high profit and suggest a combination of attributes as representing a model of high profit documents that the user could observe and redirect,
3) the computer system could present potentially “similar” documents to the set that were marked as high profit by the user,
4) the computer system could “seed” potentially high profit documents for the user to browse based on a designer-defined model of a high profit document,
5) the user could give feedback to computer-generated sets of high profit documents to “sharpen” the definition and/or “train” the computer system,
6) the computer system could remind the user to search for high profit documents during the analysis process, and
7) the computer system could critique the user’s selection of high profit documents or the reliance upon documents that are not considered high profit by the definition in the computer software.

In summary, this design seed has the characteristics of:

- helping analysts to more quickly locate high profit documents
- providing useful support to analysts even though not all high profit documents will be located and some low profit documents will be included
- visualizations and interaction mechanisms based on document attributes that can be used to characterize types of documents of interest to analysts

An example that explicitly uses the "high profit" document model to suggest candidate documents is displayed in Figure 22: the machine intelligence has a model of high profit documents that it uses to suggest candidates for the user to browse, and the user can inspect the definition of a high profit document and tailor it to a particular situation. For example, a high profit candidate for the Ariane 501 scenario used in the simulation study could be:

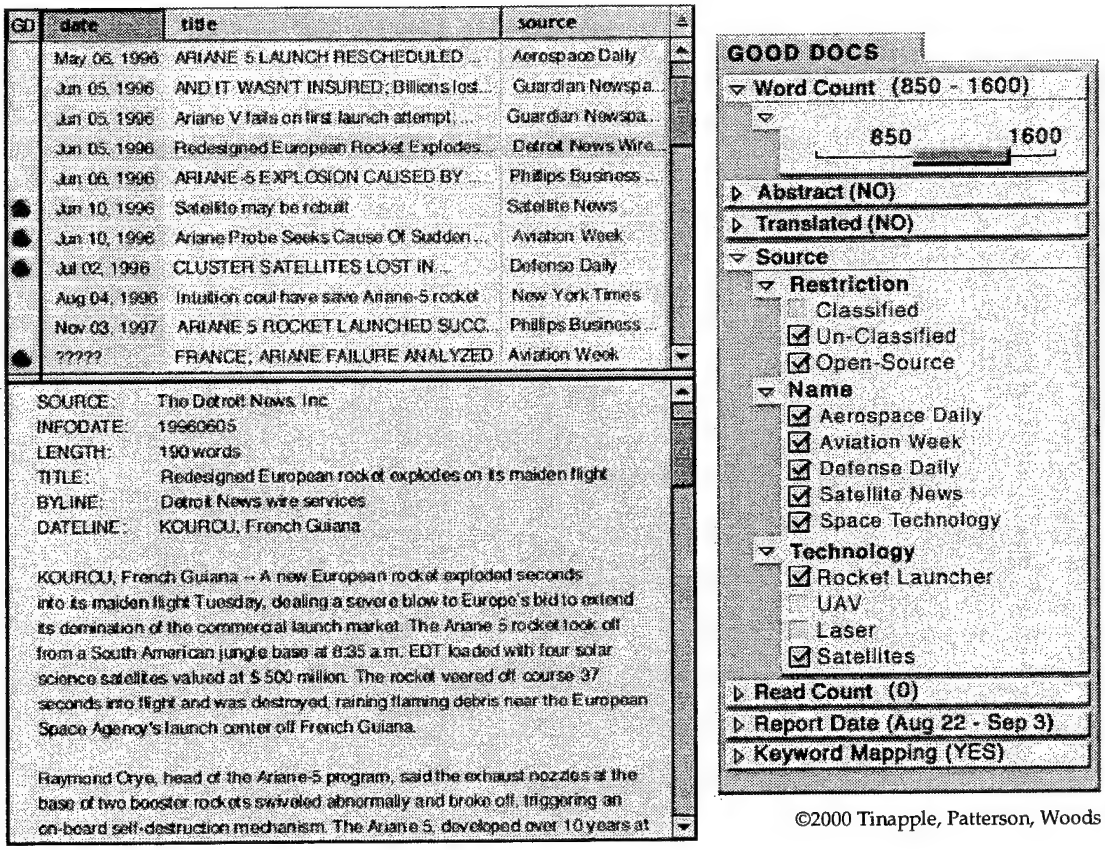
- a document with more than a thousand words,
- not an abstract,
- not translated,
- from classified, unclassified, or open sources, and
- from a pre-defined set of "good sources" for aviation tasks.



©2000 Tinapple, Patterson, Woods

Figure 22. Algorithm Identifies Candidate High Profit Documents (Query C)

A different instantiation of this design seed that might be more useful than the previous instantiation if many of the machine recommendations are incorrect is shown in Figure 23. With this version, the system is less deeply committed to an accurate model of high profit documents. A user marks what he or she considers to be good documents while browsing a set, the machine intelligence synthesizes the information and provides an inspectable set of document attributes, the user modifies the set, and then the machine uses that information to conduct a search for similar documents.



GD	date	title	source
	May 06, 1996	ARIANE 5 LAUNCH RESCHEDULED ...	Aerospace Daily
	Jun 05, 1996	AND IT WASN'T INSURED; Billions lost...	Guardian Newspa...
	Jun 05, 1996	Ariane V fails on first launch attempt; ...	Guardian Newspa...
	Jun 05, 1996	Redesigned European Rocket Explodes...	Detroit News Wire...
	Jun 06, 1996	ARIANE-5 EXPLOSION CAUSED BY ...	Philips Business ...
●	Jun 10, 1996	Satellite may be rebuilt	Satellite News
●	Jun 10, 1996	Ariane Probe Seeks Cause Of Sudden ...	Aviation Week
●	Jul 02, 1996	CLUSTER SATELLITES LOST IN ...	Defense Daily
	Aug 04, 1996	Intuition coul have save Ariane-5 rocket	New York Times
	Nov 03, 1997	ARIANE 5 ROCKET LAUNCHED SUCC...	Philips Business ...
●	?????	FRANCE; ARIANE FAILURE ANALYZED	Aviation Week

SOURCE: The Detroit News, Inc
INFODATE: 19960605
LENGTH: 190 words
TITLE: Redesigned European rocket explodes on its maiden flight
BYLINE: Detroit News wire services
DATELINE: KOUROU, French Guiana

KOUROU, French Guiana -- A new European rocket exploded seconds into its maiden flight Tuesday, dealing a severe blow to Europe's bid to extend its domination of the commercial launch market. The Ariane 5 rocket took off from a South American jungle base at 8:35 a.m. EDT loaded with four solar science satellites valued at $ 500 million. The rocket veered off course 37 seconds into flight and was destroyed, raining flaming debris near the European Space Agency's launch center off French Guiana.

Raymond Orye, head of the Ariane-5 program, said the exhaust nozzles at the base of two booster rockets swiveled abnormally and broke off, triggering an on-board self-destruction mechanism. The Ariane 5, developed over 10 years at

GOOD DOCS
- Word Count (850 - 1600)
  - 850 – 1600
- Abstract (NO)
- Translated (NO)
- Source
  - Restriction
    - □ Classified
    - ☑ Un-Classified
    - ☑ Open-Source
  - Name
    - ☑ Aerospace Daily
    - ☑ Aviation Week
    - ☑ Defense Daily
    - ☑ Satellite News
    - ☑ Space Technology
  - Technology
    - ☑ Rocket Launcher
    - □ UAV
    - □ Laser
    - ☑ Satellites
- Read Count (0)
- Report Date (Aug 22 - Sep 3)
- Keyword Mapping (YES)

©2000 Tinapple, Patterson, Woods

Figure 23. User Identifies "Good" Documents and Machine Finds Similar Documents Based on Modifiable Attributes

In the situation in Figure 23, the "GD" column refers to "good documents" and is available as a default in the interface. A possible extension to this design idea would be user-defined categories of documents other than "good documents." For example, the user might want to group and search for documents that speak to a particular sub-theme such as how insurance rates would be affected by the Ariane 501 launch failure. In addition to the "GD" column, other types of documents could be identified, such as the ones described in Table 11.

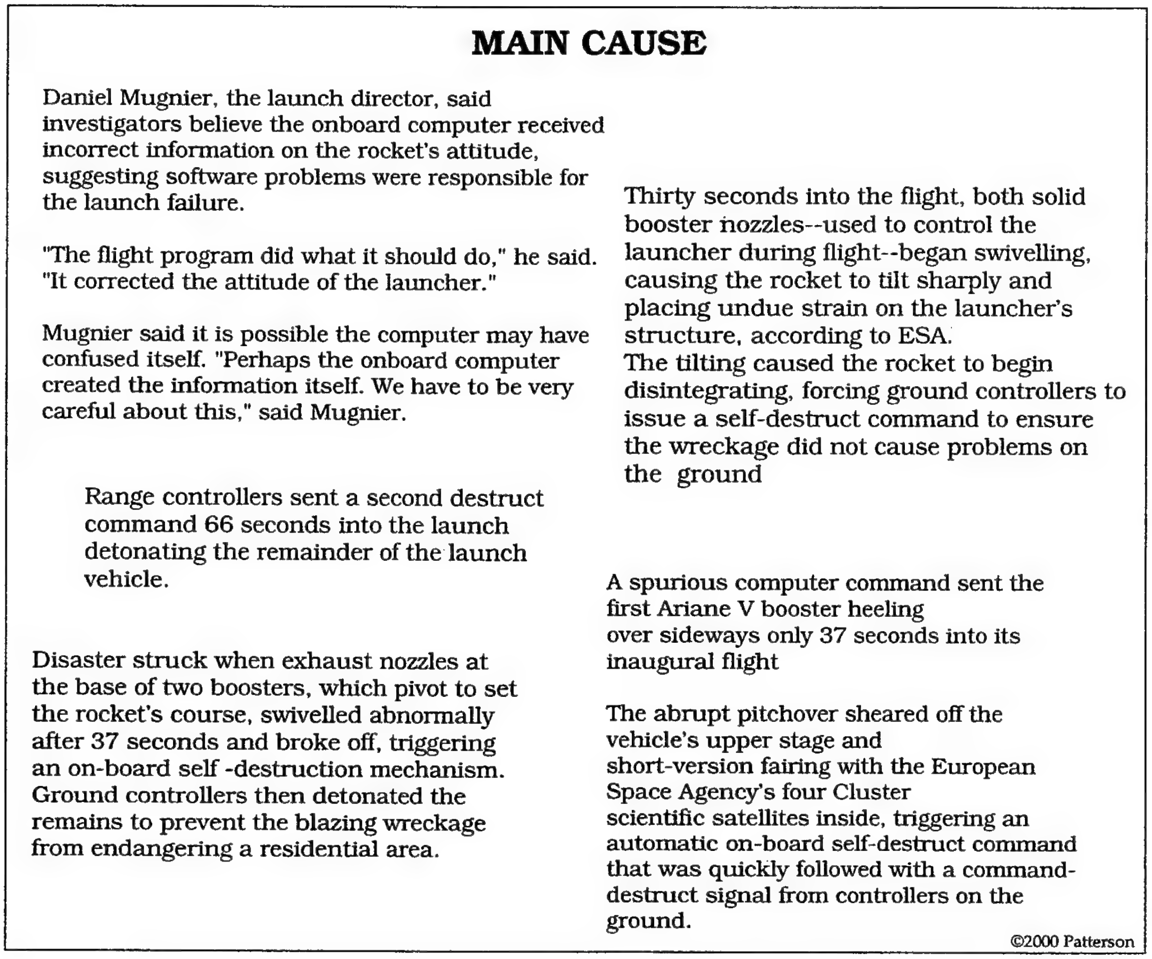
Table 11. Potential Document Categories

Type of Document	Description of Category
High Profit	Detailed, accurate description of important events from a credible, low bias source after most of the information about the event has become available
On Topic	Contains information that is relevant to the analyst's task
Comprehensive	Long, on topic article from a credible, low-bias source that is not immediately after the event and not overly distanced from the original data
Peripheral Mentions	Documents that reference an event or topic briefly but whose focus is on themes of low relevance to the user's tasks and goals
Unrelated Collections	Documents that contain many unrelated items that are not ordered in a meaningful way (e.g., "Briefs")
Themed	Documents that address a particular theme
Post-Announcements	Documents that announce a past event or communication but without many other details or analysis of the impacts
Pre-Announcements	Documents that announce a planned event or communication but without many other details or analysis of the potential impacts

4.8 Design Seed 4: Manipulating Text "Snippets"

Many of the study participants were seen to take small portions of opened documents into a word processing program. Comments by several of the study participants indicated that support tools were targeted at helping to search for information, but that virtually no tools existed for what they referred to as the "analysis" part of the process. This "analysis" generally referred to the difficult cognitive work involved in breaking down documents into thematic components, improving and assessing the quality of the information available on those components, and reassembling the information into a coherent story that could be defended.

It became clear from observing the study participants work with these selected portions of documents, or “snippets,” that a substantial amount of data manipulation was performed at this data unit (Figures 24 and 25 show “snippets” from 20 documents on themes from the Ariane 501 scenario).



MAIN CAUSE

Daniel Mugnier, the launch director, said investigators believe the onboard computer received incorrect information on the rocket's attitude, suggesting software problems were responsible for the launch failure.

"The flight program did what it should do," he said. "It corrected the attitude of the launcher."

Mugnier said it is possible the computer may have confused itself. "Perhaps the onboard computer created the information itself. We have to be very careful about this," said Mugnier.

Range controllers sent a second destruct command 66 seconds into the launch detonating the remainder of the launch vehicle.

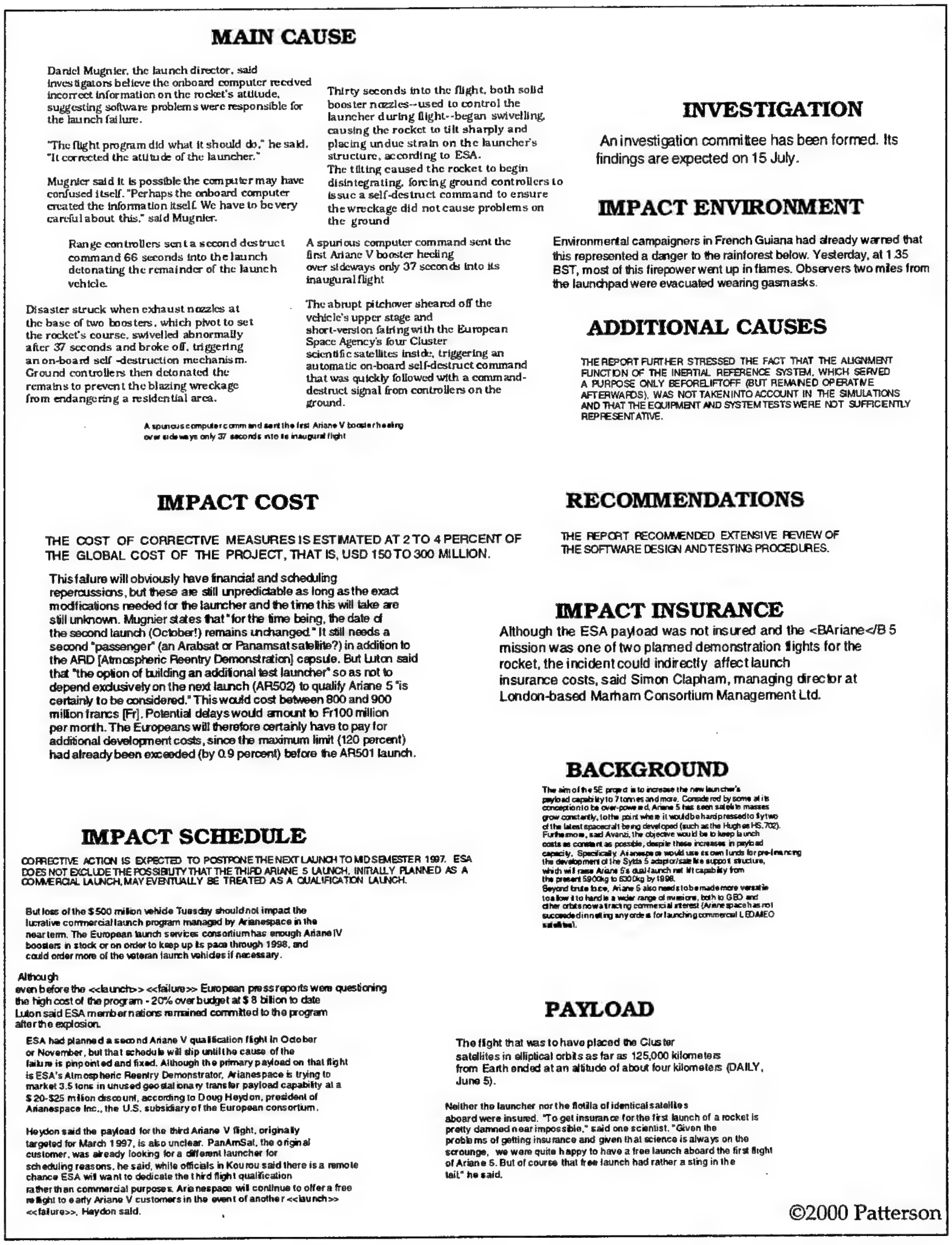
Disaster struck when exhaust nozzles at the base of two boosters, which pivot to set the rocket's course, swivelled abnormally after 37 seconds and broke off, triggering an on-board self -destruction mechanism. Ground controllers then detonated the remains to prevent the blazing wreckage from endangering a residential area.

Thirty seconds into the flight, both solid booster nozzles--used to control the launcher during flight--began swivelling, causing the rocket to tilt sharply and placing undue strain on the launcher's structure, according to ESA.
The tilting caused the rocket to begin disintegrating, forcing ground controllers to issue a self-destruct command to ensure the wreckage did not cause problems on the ground

A spurious computer command sent the first Ariane V booster heeling over sideways only 37 seconds into its inaugural flight

The abrupt pitchover sheared off the vehicle's upper stage and short-version fairing with the European Space Agency's four Cluster scientific satellites inside, triggering an automatic on-board self-destruct command that was quickly followed with a command-destruct signal from controllers on the ground.

Figure 24. Snippets about the Main Cause of the Ariane 501 Explosion



MAIN CAUSE

Daniel Mugnier, the launch director, said investigators believe the onboard computer received incorrect information on the rocket's attitude, suggesting software problems were responsible for the launch failure.

"The flight program did what it should do," he said. "It corrected the attitude of the launcher."

Mugnier said it is possible the computer may have confused itself. "Perhaps the onboard computer created the information itself. We have to be very careful about this," said Mugnier.

Thirty seconds into the flight, both solid booster nozzles--used to control the launcher during flight--began swivelling, causing the rocket to tilt sharply and placing undue strain on the launcher's structure, according to ESA.
The tilting caused the rocket to begin disintegrating, forcing ground controllers to issue a self-destruct command to ensure the wreckage did not cause problems on the ground

Range controllers sent a second destruct command 66 seconds into the launch detonating the remainder of the launch vehicle.

A spurious computer command sent the first Ariane V booster heeling over sideways only 37 seconds into its inaugural flight

Disaster struck when exhaust nozzles at the base of two boosters, which pivot to set the rocket's course, swivelled abnormally after 37 seconds and broke off, triggering an on-board self -destruction mechanism. Ground controllers then detonated the remains to prevent the blazing wreckage from endangering a residential area.

The abrupt pitchover sheared off the vehicle's upper stage and short-version fairing with the European Space Agency's four Cluster scientific satellites inside, triggering an automatic on-board self-destruct command that was quickly followed with a command-destruct signal from controllers on the ground.

A spurious computer command sent the first Ariane V booster heeling over sideways only 37 seconds into its inaugural flight

INVESTIGATION

An investigation committee has been formed. Its findings are expected on 15 July.

IMPACT ENVIRONMENT

Environmental campaigners in French Guiana had already warned that this represented a danger to the rainforest below. Yesterday, at 1.35 BST, most of this firepower went up in flames. Observers two miles from the launchpad were evacuated wearing gasmasks.

ADDITIONAL CAUSES

THE REPORT FURTHER STRESSED THE FACT THAT THE ALIGNMENT FUNCTION OF THE INERTIAL REFERENCE SYSTEM, WHICH SERVED A PURPOSE ONLY BEFORELIFTOFF (BUT REMAINED OPERATIVE AFTERWARDS), WAS NOT TAKEN INTO ACCOUNT IN THE SIMULATIONS AND THAT THE EQUIPMENT AND SYSTEM TESTS WERE NOT SUFFICENTLY REPRESENTATIVE.

IMPACT COST

THE COST OF CORRECTIVE MEASURES IS ESTIMATED AT 2 TO 4 PERCENT OF THE GLOBAL COST OF THE PROJECT, THAT IS, USD 150 TO 300 MILLION.

This failure will obviously have financial and scheduling repercussions, but these are still unpredictable as long as the exact modifications needed for the launcher and the time this will take are still unknown. Mugnier states that "for the time being, the date of the second launch (October!) remains unchanged." It still needs a second "passenger" (an Arabsat or Panamsat satellite?) in addition to the ARD [Atmospheric Reentry Demonstration] capsule. But Luton said that "the option of building an additional test launcher" so as not to depend exclusively on the next launch (AR502) to qualify Ariane 5 "is certainly to be considered." This would cost between 800 and 900 million francs [Fr]. Potential delays would amount to Fr100 million per month. The Europeans will therefore certainly have to pay for additional development costs, since the maximum limit (120 percent) had already been exceeded (by 0.9 percent) before the AR501 launch.

RECOMMENDATIONS

THE REPORT RECOMMENDED EXTENSIVE REVIEW OF THE SOFTWARE DESIGN AND TESTING PROCEDURES.

IMPACT INSURANCE

Although the ESA payload was not insured and the <BAriane</B 5 mission was one of two planned demonstration flights for the rocket, the incident could indirectly affect launch insurance costs, said Simon Clapham, managing director at London-based Marham Consortium Management Ltd.

BACKGROUND

The aim of the 5E project is to increase the new launcher's payload capability to 7 tonnes and more. Considered by some at its conception to be over-powered, Ariane 5 has seen satellite masses grow constantly, to the point where it would be hard pressed to fly two of the latest spacecraft being developed (such as the Hughes HS.702). Furthermore, said Avanzi, the objective would be to keep launch costs as constant as possible, despite these increases in payload capacity. Specifically, Arianespace would use its own funds for pre-financing the development of the Sylda 5 adaptor/satellite support structure, which will raise Ariane 5's dual-launch net lift capability from the present 5900kg to 6300kg by 1998.
Beyond brute force, Ariane 5 also needs to be made more versatile to allow it to handle a wider range of missions, both to GEO and other orbits now attracting commercial interest (Arianespace has not succeeded in netting any orders for launching commercial LEO/MEO satellites).

IMPACT SCHEDULE

CORRECTIVE ACTION IS EXPECTED TO POSTPONE THE NEXT LAUNCH TO MID SEMESTER 1997. ESA DOES NOT EXCLUDE THE POSSIBILITY THAT THE THIRD ARIANE 5 LAUNCH, INITIALLY PLANNED AS A COMMERCIAL LAUNCH, MAY EVENTUALLY BE TREATED AS A QUALIFICATION LAUNCH.

But loss of the $500 million vehicle Tuesday should not impact the lucrative commercial launch program managed by Arianespace in the near term. The European launch services consortium has enough Ariane IV boosters in stock or on order to keep up its pace through 1998, and could order more of the veteran launch vehicles if necessary.

Although
even before the <<launch>> <<failure>> European press reports were questioning the high cost of the program - 20% over budget at $ 8 billion to date Luton said ESA member nations remained committed to the program after the explosion.

ESA had planned a second Ariane V qualification flight in October or November, but that schedule will slip until the cause of the failure is pinpointed and fixed. Although the primary payload on that flight is ESA's Atmospheric Reentry Demonstrator, Arianespace is trying to market 3.5 tons in unused geostationary transfer payload capability at a $ 20-$25 million discount, according to Doug Heydon, president of Arianespace Inc., the U.S. subsidiary of the European consortium.

Heydon said the payload for the third Ariane V flight, originally targeted for March 1997, is also unclear. PanAmSat, the original customer, was already looking for a different launcher for scheduling reasons, he said, while officials in Kourou said there is a remote chance ESA will want to dedicate the third flight qualification rather than commercial purposes. Arianespace will continue to offer a free reflight to early Ariane V customers in the event of another <<launch>> <<failure>>, Haydon said.

PAYLOAD

The flight that was to have placed the Cluster satellites in elliptical orbits as far as 125,000 kilometers from Earth ended at an altitude of about four kilometers (DAILY, June 5).

Neither the launcher nor the flotilla of identical satellites aboard were insured. "To get insurance for the first launch of a rocket is pretty damned near impossible," said one scientist. "Given the problems of getting insurance and given that science is always on the scrounge, we were quite happy to have a free launch aboard the first flight of Ariane 5. But of course that free launch had rather a sting in the tail," he said.

©2000 Patterson

Figure 25. Snippets Grouped by Themes in the Ariane 501 Scenario

Support is clearly needed for aiding in the manipulation of snippets, analogous to tacking and moving index cards on a bulletin board or writing on a whiteboard. As analysts read documents, they need to be able to place portions of text into a work area and the information about where the text had originated should be automatically available on demand (such as by a hyperlink). Often after most of the information is gathered together, they need support for grouping, labeling, and tagging items.

In summary, this design seed has the characteristics of:

- allowing analysts to place selected snippets in a dedicated work area,
- viewing information from different documents in parallel,
- "remembering" where snippets originated, and
- supporting grouping, labeling, and tagging snippets.

As a byproduct of having the information in an electronic format, there are opportunities for intelligent machine processing to go beyond this baseline support. For example, machine processing might potentially:

- group together descriptions on the same topic,
- identify data conflicts and corroborations across descriptions,
- remind analysts to verify information in a snippet before incorporating it into an analytic product,
- look for similar snippets in the database,
- track who annotates information during collaborative analyses,
- sort snippets by different attributes (e.g., date, source, length, topic), and
- do "what if" simulations on particular sources to see the vulnerability of the overall analysis to a poor source.

Note that the interface design for this design seed is comparatively easy because the majority of the interface is organized by how the analyst groups the snippets. Nevertheless, it is interesting to point out that creating the Animock for this design seed within a scenario identified several usability challenges that might not have been confronted otherwise. For example, it was quickly realized that there would not be enough interface "real estate" to read all of the snippets in parallel. We considered using "windowshades," tabs, or hyperlinks to increase the available space. However, we quickly realized that these navigation features would require extensive screen management and reduce some of the benefit of the support tool. Then we investigated some existing software packages that displayed large amounts of information. We discovered that natural zooming interactions, coupled with "longshot" labels that are always available at the most zoomed out view, could greatly increase the amount of information that could be displayed and manipulated at the snippet level in one window.

4.9 Design Seed 5: Conflict and Corroboration Support

One of the most important requirements of an analytic product is that it be accurate, or at least that the analyst be well-calibrated as to its accuracy. This is one of the main reasons that the same report written by an intelligence analyst and an unknown author are viewed differently: by putting an intelligence analysis organization's name on the report, there is an implicit "stamp of approval" that increases its value.

Study participants clearly viewed the elimination of inaccuracies by finding converging evidence across independent sources as a major component of the value of an analytic product. The participants described and employed a variety of strategies for tracking and resolving discrepant descriptions in order to reduce their vulnerability to incorporating inaccurate information. Partly because this cognitively difficult process of corroborating information and resolving conflicting information was unsupported by the tools that they were provided, nearly every participant experienced some breakdowns in this process.

During the study, the participants described strategies that they would use if they had more time to conduct an analysis to ensure that information was corroborated, such as printing out the documents and highlighting topics with particular colors every time they appeared from independent sources. However, in general, no study participants used these strategies during the simulated task because of the short deadline, and described that under high workload conditions they tended to shed this task in the workplace as well. In addition, several noted that their strategies were not well-supported within the electronic environment (e.g., it is difficult to see information from multiple documents in parallel on the screen). Supporting these strategies would not be overly difficult in an electronic environment, and this support forms the essence of this design seed.

In summary, this design seed has characteristics of supporting the following tasks to improve the accuracy of an analytic product:

- identifying data conflicts,
- highlighting uncertain information,
- remembering judgments about relationships between data,
- tracking "loose ends" that need to be resolved later,
- identifying when data comes from the same original source[7], and
- identifying attempts of others to purposely misinform and deceive.

The primary design challenge with this design seed is how to minimize data entry while still supporting the observed strategies. Although there are many theoretically possible categories of relationships among data elements (Schum, 1994), in the strategies we observed, analysts only documented broad-brush distinctions, e.g., same topic, uncertain, verified. Similarly, they did not explicitly say what information was conflicting, just that they previously saw something that they thought contradicted what they were reading then. Therefore, design seeds that require overly specific data relationship categories

[7] Note that intelligence analysts refer to a potential vulnerability in analysis as "creeping validity." This phrase is used to refer to situations where multiple reports appear to corroborate an event or other piece of information that actually came from the same original source. In these cases, even though there are multiple accounts, belief in the accuracy of the account should not be increased.

or require specifying what information conflicts with each other might be judged to require too much work on the user's part to be useful.

As with the other design seeds, the level of machine intelligence could vary greatly based on how much information is available to the machine intelligence to process and how often the machine processing will be incorrect. At the lowest level of machine intelligence, the software could simply display judgments made by intelligence analysts, such as by displaying an underline mark under a word that an analyst selected to be underlined for an unknown reason. At the highest end, a user could ask the machine to critique the process that had been followed in verifying the information was accurate. The machine could then:

- analyze the breadth of information sampling in time and in high profit documents,
- identify potential data conflicts both in the information that was read as well as potentially available in databases,
- check that information that was marked as corroborating came from independent sources, and
- assess the quality of the documents that were most heavily used in the analysis (most likely based on what information has been pulled or marked from what documents).

4.10 Design Seed 6: Timeline Construction Aid

Although much of the analytic work conducted by the study participants was done "in the head" of the analysts without any support tools (including pencil and paper), there were several occasions where they described the need to plot events on a timeline. One of the study participants did this in order to determine which rocket launch failure was the one of interest in the Quick Reaction Task. At this timescale, relevant other failures would include what is displayed in Figure 26.

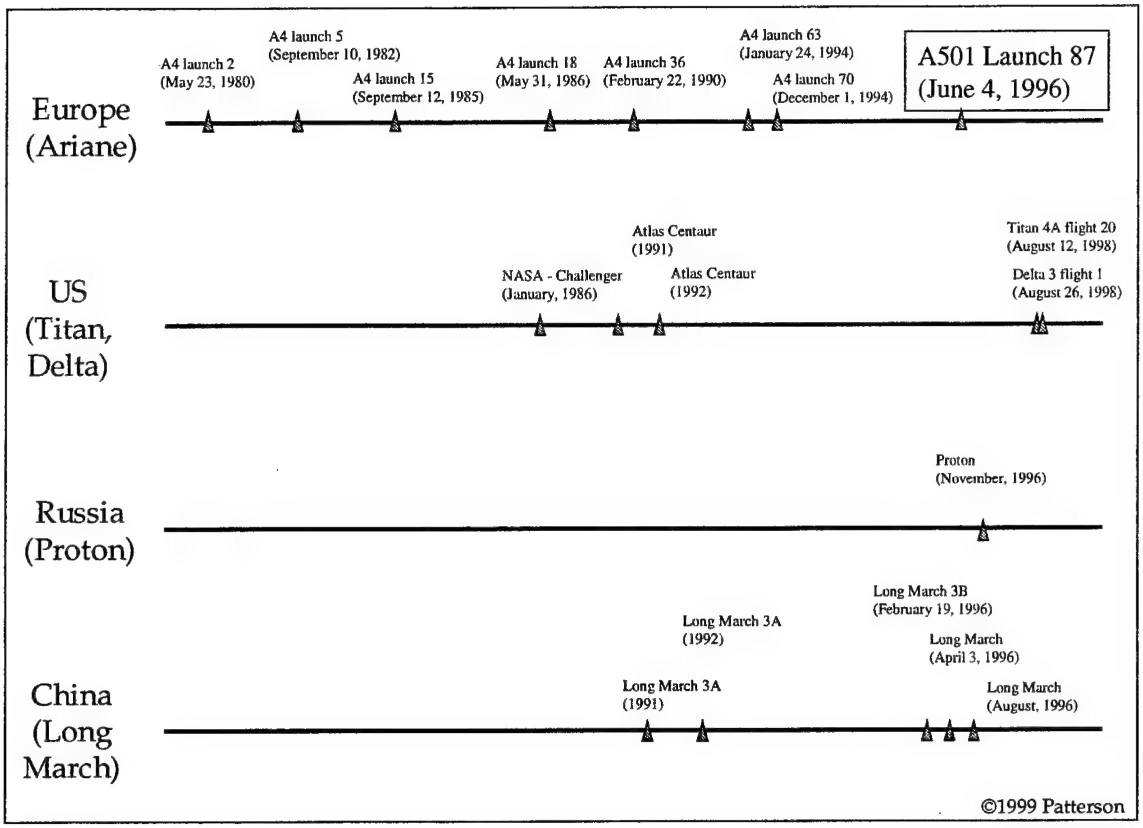


Figure 26. Rocket Launch Failures in Satellite Launch Industry

At a much smaller timescale (seconds instead of years), study participants were concerned about the sequence of events that occurred during the launch timeline. Figure 27 shows the sequence of events that occurred in the first minute of the Ariane 501 launch.

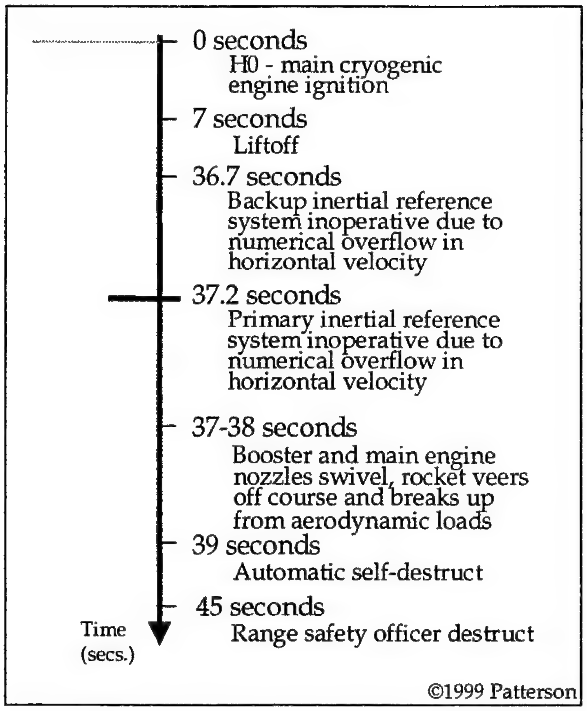
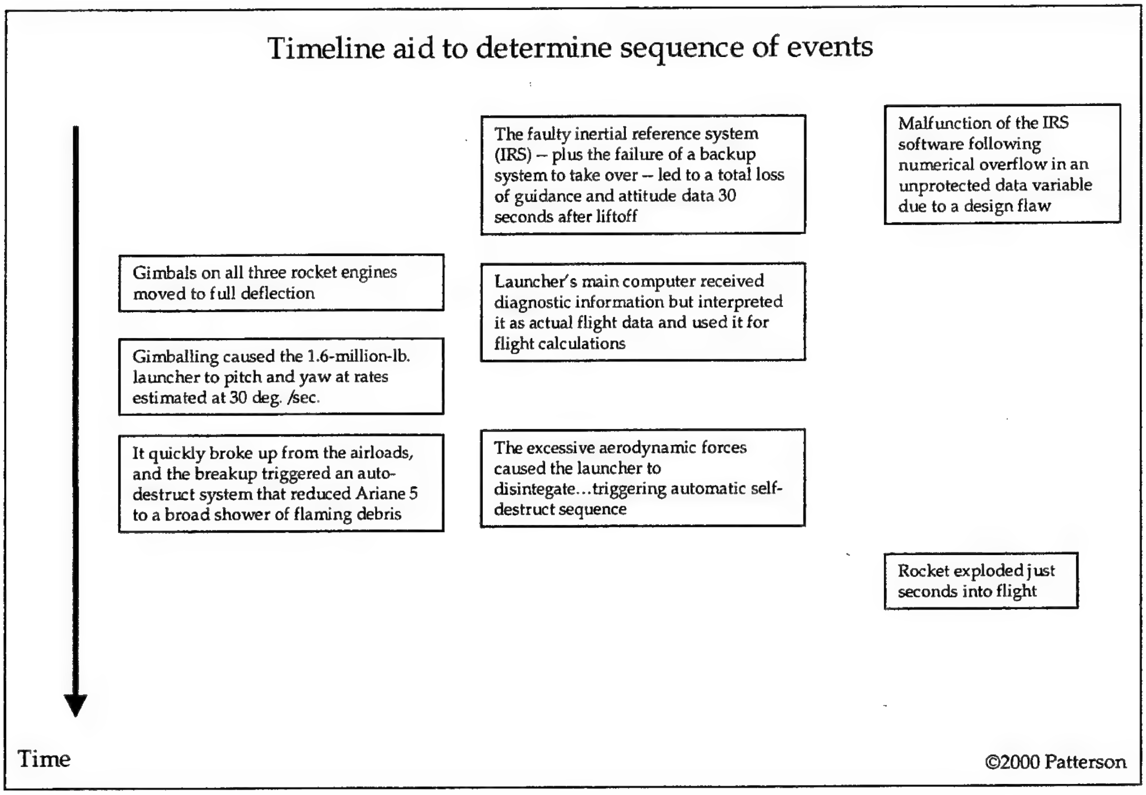

0 seconds
H0 - main cryogenic engine ignition
7 seconds
Liftoff
36.7 seconds
Backup inertial reference system inoperative due to numerical overflow in horizontal velocity
37.2 seconds
Primary inertial reference system inoperative due to numerical overflow in horizontal velocity
37-38 seconds
Booster and main engine nozzles swivel, rocket veers off course and breaks up from aerodynamic loads
39 seconds
Automatic self-destruct
45 seconds
Range safety officer destruct
Time (secs.)
©1999 Patterson

Figure 27. Sequence of Events During Ariane 501 Take-off

This design seed is an idea of supporting the analytic work that results in figures such as Figures 26 and 27. Note, however, that this design seed is not necessarily intended to support the creation of figures like these that would go directly into a briefing slide or written report. Rather, the concept is to support the analytic work that precedes the final documentation. This distinction between these types of support is made because these representations are constructed bottom-up over time and we do not wish to require users to enter data such as a specific time in order to plot events in sequence. Instead, we envision something more like what is displayed in Figure 28. Users can copy and paste or create "snippets" in a workspace that is organized by time and allows implicit "threads" to emerge over time through user-determined spacing alignment. As the analysis progresses, users can become more explicit about particular times events occur and label themes in ways that lead towards a figure that can be exported to an analytic product.

In summary, this design seed has the characteristics of:

- allowing themes and event sequences to emerge from spatial manipulation of snippets
- supporting timeline analyses at multiple timescales.



Timeline aid to determine sequence of events

Gimbals on all three rocket engines moved to full deflection

Gimballing caused the 1.6-million-lb. launcher to pitch and yaw at rates estimated at 30 deg. /sec.

It quickly broke up from the airloads, and the breakup triggered an auto-destruct system that reduced Ariane 5 to a broad shower of flaming debris

The faulty inertial reference system (IRS) -- plus the failure of a backup system to take over -- led to a total loss of guidance and attitude data 30 seconds after liftoff

Launcher's main computer received diagnostic information but interpreted it as actual flight data and used it for flight calculations

The excessive aerodynamic forces caused the launcher to disintegate...triggering automatic self-destruct sequence

Malfunction of the IRS software following numerical overflow in an unprotected data variable due to a design flaw

Rocket exploded just seconds into flight

©2000 Patterson

Figure 28. Timeline Aid for Organizing Snippets about Events

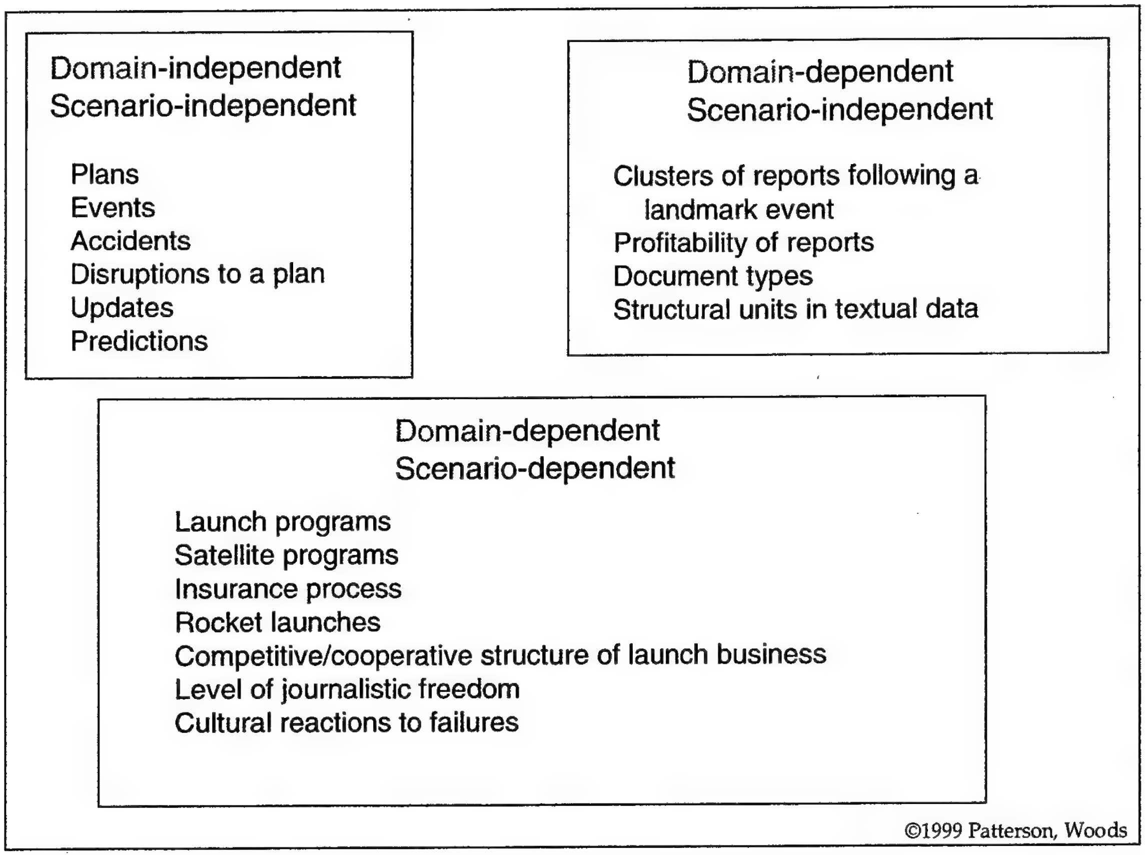
This design seed could potentially take advantage of advances in natural language processing to recognize events that might need to be incorporated on a timeline. Whether or not these machine-recognized events would be directly incorporated into an interface with snippets pulled out by an analyst would depend on how many events are recognized. Some current Natural Language Understanding (NLU) processors recognize events in every sentence. Clearly, this many machine-recognized events would dominate a display that combined human-recognized and machine-recognized events. On the other hand, if heuristics could be applied such as requiring a flurry of reports within a short time period for documents containing a particular combination of keywords and attributes, then there would be much fewer machine-suggested events. Similarly, if machine processing was limited to recognizing events within a select group of documents or at a certain timescale, the NLU processors might prove useful in "seeding" a timeline representation, tracking events, and looking for similar events within a set.

4.11 Design Seed 7: Using Context-Specific Models to "Seed" Themes

Our "diagnosis" of data overload (Woods et al., 1998) found that an extremely difficult challenge when addressing the data overload problem is that data is only informative given a particular context. Context sensitivity plays a central role in many of our design seeds to address data overload, and makes many of our ideas distinct from current design directions. We believe that using the basic human competence for finding what is informative in natural perceptual fields despite context sensitivity is our guide for innovation. With this approach, our goals are to use the power of technology:

- to enhance observability,
- to take into account context sensitivity, and
- to build conceptual spaces.

One way to take into account context sensitivity is to use the semantics of underlying processes or field of activity to help define the relationships that give data meaning (Vicente and Rasmussen, 1992). For different analytic scenarios, there will be multiple organizing themes, each of which defines a perspective on the data field. In the Ariane 501 scenario, there were many potential models at different levels of abstraction (Figure 29) that could be leveraged in a context-sensitive approach, including models independent of both the satellite industry and the specific scenario that could be used in a variety of analysis tasks.



©1999 Patterson, Woods

Figure 29. Models in Ariane 501 Scenario

There are a variety of ways that machine intelligence could take advantage of models such as these to aid analysts during the analysis process. One strategy would be for the computer to "seed" a display with initial themes to consider pursuing in the analysis. These themes could be available to the machine from a variety of sources:

- a "modelbase" designed into the software
- themes from past analyses that are recognized as similar such as with case-based recognition algorithms
- Natural Language Understanding (NLU) processing of a written or verbal question
- A tailored "modelbase" created from a machine synthesis of past analyses or crafted by an individual or team of analysts.

In addition to seeding potential themes, machine intelligence could be used to "know" something about themes that helps with their management. For example, themes can be suggested to be added or removed based on particular heuristics. Similarly, labels for themes can be suggested from analysis of documents that are "attached" to themes. Additionally, the machine could be asked to organize themes based on heuristics about how certain themes relate to each other (e.g., "background" comes first and "impacts" goes last).

In summary, this design seed has characteristics of:

- Using context-specific knowledge about a domain and/or scenario to guide searches for information and organization of data
- Encouraging a meta-analysis of themes to include in an analytic product

Benefits to this design seed on performance would hopefully include encouraging a "meta-analysis" of what themes to include in an analysis task before beginning. These themes can then be used to better guide the search for information and the determination of when an analysis should stop. By determining what to look for in advance, hopefully it would be more difficult for an analyst to get sidetracked onto tangential themes or to let personal preferences about interesting topics dominate an analytic product.

As an illustration of how the approach of instantiating design seeds in Animocks can lead to discussions at the "usefulness" level of design as opposed to "usability" comments, such as feedback about color choices on a display, consider the reaction of an expert analyst:

> "So, the problem, as I see it here is finding a way to identify a "theme" and identifying what is significant to that theme and somehow associating it with the theme. Now, if I have 100,000 documents and identify 100 documents in some theme process; I notice that I have only got 0.1% of the pile. This may be good; or this may be bad. I can probably work with only 100 documents. But the question remains: what is still in the other 99.9%? Is there anything there that has a bearing on what I know from the 0.1%? How can I best satisfy myself that the rest is all trash for this exercise (I might hurriedly do a list of titles and scan them for whether something will catch my eye or not)?"

This feedback is very important because we can then add another desirable characteristic of this design seed: that a dataset be characterized by themes so that the analyst can verify that (s)he has a sense of what themes are available and what the possibilities are that (s)he has missed a critical theme. As further evidence that we are engaging in a fruitful discussion of what would be useful to design in order to reduce the risk of designing a system that will be later underutilized or rejected by end users, the analyst further notes that "there may be other ways to achieve [this goal]" after describing a particular approach that could be taken:

> "Another approach to the above would be to have a large set of "canned" themes and have them run against the same set of data simultaneously...if the display shows a fabric representing the

> 100K documents with a bunch of "theme" peaks all over the topography that I can access to see what they represent. Then I feel pretty good when I find only one peak on my primary interest and all the others are so far removed from my interest that I can disregard them."

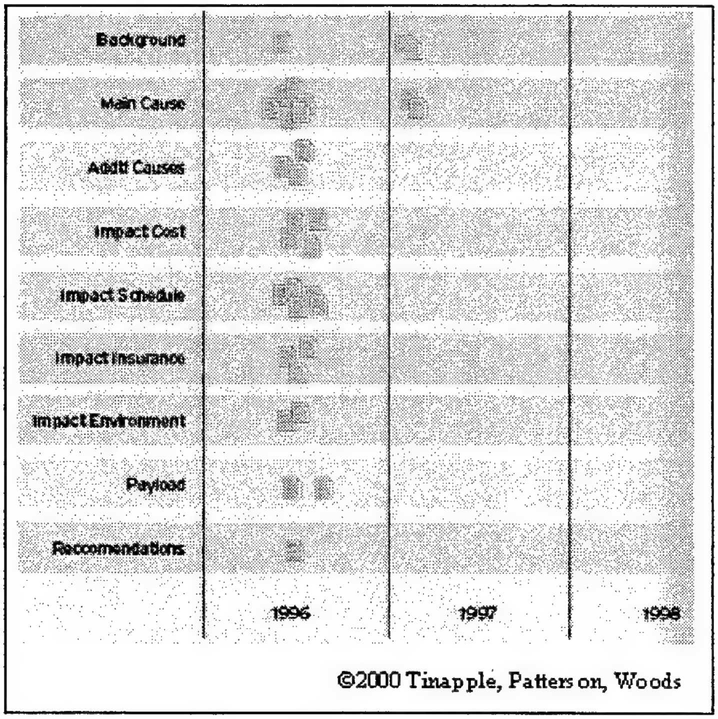
Note that any model-based method to depict more than base data is subject to the "right" model catch -- how do you know the model that specifies how data is informative is appropriate for the task or situation? Also, as was pointed out by the expert analyst, "The "canned theme" might have to be regularly updated to meet changing times (that could be a "downer")." Although we do not have any completely satisfying solution to this catch, we advise that the human user always be allowed to override the suggestions of a machine processor as well as update or otherwise alter the available "modelbase." In order to make a useful and usable design concept based on this design seed, we would want to explore further how to address these concerns. For, as the analyst notes, "The problem is immensely complex."

4.12 Design Seed 8: Longshot Themespace Visualization

Whether or not machine intelligence is used to "seed" themes based on pre-defined models as described in the previous design seed, we feel that it is important to support analysts in stepping back and thinking critically about the analytic process that had been followed. Although several software packages support creating notes at the "snippet" level, we have seen no software with explicit support for stepping back to see gaps in an analysis process.

A visualization that gives the overall view of snippets organized by themes (Figure 30) is intended to have the characteristics of:

- Supporting meta-level review of the analysis process that has been used and the information that has been sampled (i.e., "seeing the forest for the trees"),
- Making gaps in analysis salient,
- Making "loose ends" to be remembered salient,
- Supporting reorganization of the "snippet" details by manipulation of higher order abstractions such as hierarchical and sequential relationships of themes,
- Supporting development of a narrative, including sequence of landmark events and critical assessments of purposeful activity, and
- Supporting principled judgments of when to stop an analysis process.



©2000 Tinapple, Patterson, Woods

Figure 30. Longshot View of Snippets in Themespace

In designing the Animock for this design seed, we identified several usability-level issues that would need to be explicitly addressed in the design of an actual system. The visualization in Figure 30 could potentially represent any of the following levels of textual data: reports that were browsed, reports that were opened, reports that were marked in some fashion, portions of reports, "snippets" that were copied to a workspace, snippets that were marked in some way such as placed in a spatially dedicated area to represent inclusion on the visualization, or notes that were written by an analyst. Which of these data would be represented on the visualization would need to be decided based on how much work would be required on the analyst's part to manipulate "snippets" and label snippet groupings. It is unlikely that support tools that require labels or explicit decisions about how snippets relate to themes as they are moved into a workspace will be viewed as useful as tools that allow themes and groupings to naturally emerge over time.

In addition to the baseline support that we expect that a longshot visualization of themes would provide, there are several extensions to this concept that might prove useful. For example, the themespace concept could be more explicitly used to support returning to the available data for targeted searches to fill either human-identified or machine-identified gaps in the analysis. Also, the themespace visualization could be used to aid in cooperative work among analysts. For example, analysts directly working together could be supported in discussions of how to divide analytic work by the themespace representation. Alternatively, the themespace representation could be used to provide further information during critiques of the analytic process than a written briefing. Finally, themespace representations could be used to support completing an analysis process that has already been started, either by the same analyst some time later or by a different analyst who takes over the analytic responsibility.

4.13 Design Seed 9: Finding Updates

The final design seed is an attempt to address what is probably the most difficult challenge in time-pressured computer-supported inferential analysis under data overload conditions. When analyzing the data from the study participants, a surprising finding was that the study participant who had the most prior knowledge of the Ariane 501 scenario, the most technical knowledge about rocket launcher technology, spent the most time during the analysis, and generated a written briefing in addition to a verbal briefing made an inaccurate statement in the written briefing. This inaccurate statement appears to be explained by a particularly difficult challenge during data analysis: detecting updates to once-believed-accurate information. Because the "findings" or data set on which the analysis was based came in over time, there was always the possibility of missing information that was released after the report that was being read that could overturn or render previous information "stale" (see Figure 31 for examples in the Ariane 501 scenario). When these updates occurred on themes that were not central enough to be included in report titles or newsworthy enough to generate a flurry of reports, it was extremely difficult to know if updates had occurred or where to look for them.

This design seed has the characteristic of:

- Helping analysts to locate updates that overturn or substantially change an analytic conclusion
- Helping analysts to calibrate their assessment of analytic accuracy to the likelihood that updates that render analytic conclusions inaccurate do not exist.

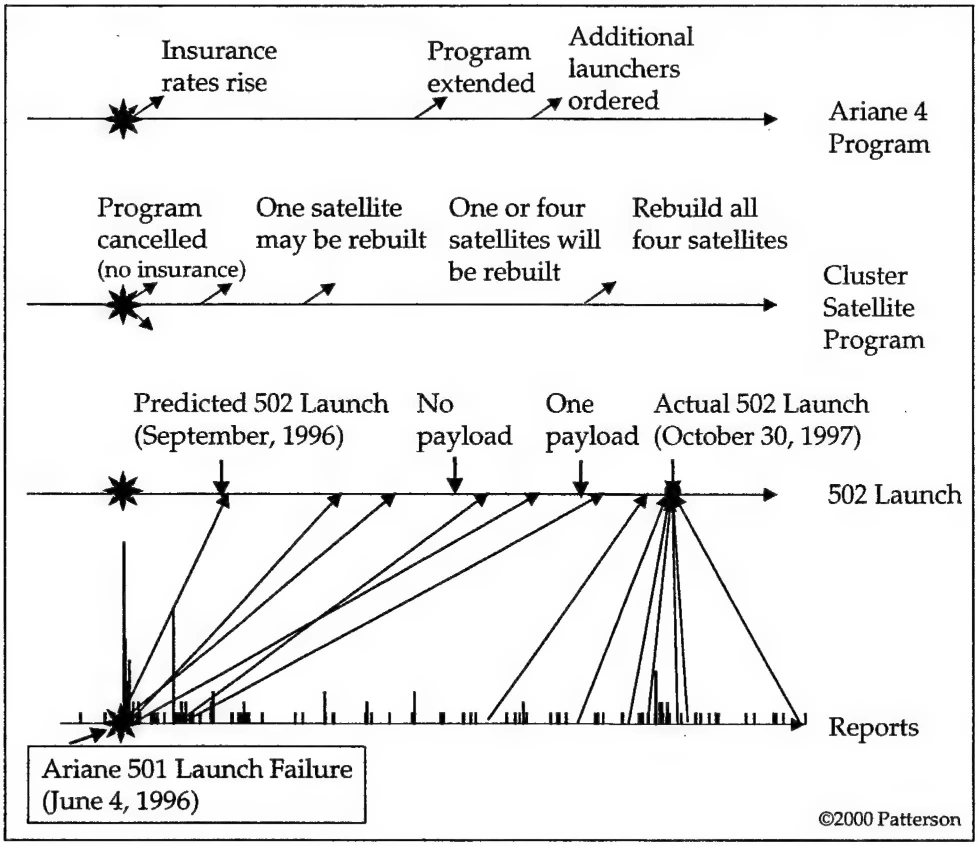

Insurance rates rise
Program extended
Additional launchers ordered
Ariane 4 Program
Program cancelled (no insurance)
One satellite may be rebuilt
One or four satellites will be rebuilt
Rebuild all four satellites
Cluster Satellite Program
Predicted 502 Launch (September, 1996)
No payload
One payload
Actual 502 Launch (October 30, 1997)
502 Launch
Reports
Ariane 501 Launch Failure (June 4, 1996)
©2000 Patterson

Figure 31. Updates that Overturned Previous Information in Ariane 501 Scenario

It is interesting to note that most of the study participants never specifically looked for updates during the analysis process or described strategies that would do so. It is possible that training analysts about the need to search for updates might be useful, although the reaction of one novice analyst to the critique that he should look for updates was that it would be very hard to do with the tools available to him. Updates could be reported hours, days, weeks, months, years, or decades after an event. Many of the updates on more minor themes in the Ariane 501 scenario did not cause a flurry of reports and were not reflected in the date/title view of the reports.

It is possible that "agents" that suggest targeted query formulations and/or "seed" representations with updates on a theme might be useful, particularly if the agents have advanced natural language processing capabilities. We believe that this design direction will require artifact-based investigations in order to gain a better understanding of how to make the concept useful. Finding updates over time is a difficult challenge for human intelligence with current tools, and yet it is likely also to be so challenging for machine intelligence that conclusions by the machine intelligence will likely be incorrect much of the time. How much can the vulnerability to missing updates be reduced simply by having the machine intelligence remind the human partner to look for updates? Are there cues to informative areas where updates might be found, such as a flurry of setpoint crossings in a short amount of time on interrelated systems? Should the machine intelligence suggest possible candidate updates, either by "seeding" a visualization or by requiring the human to explicitly consider recommended items? What advancements in machine intelligence are required to make more accurate seeding recommendations?

In addition to significant work required to ensure usefulness of this design seed, there will be substantial usability problems to address. Visualizations will need to be investigated to ensure that machine processing is observable and directable by the user in order to make the human-machine teamwork effective and avoid situations where the human agent is surprised by actions of the machine agent.

4.14 Integration of Design Seeds Through Overarching Concept: Visual Narratives

Although the design seeds have been designed to be modular so that they can be incorporated into both short and long-term development projects, we coordinated the design seeds together in order to demonstrate them as a unified Animock of an analyst working through the Ariane 501 scenario. Coordination of the design seeds is an effortful activity because it requires explicitly considering what information needs to be viewed in parallel, how changes in one part of the data space impact other data spaces, and how to enable navigation and interaction with the data in natural ways, such as via direct manipulation.

Our workspace design is integrated by an overarching concept, which we refer to as "Visual Narratives" (Figure 32). The concept is founded on the interaction of time in "report space" and "theme space." The report space shows a histogram of reports by date, which are selected by some mechanism such as keyword queries. Flurries of reports in time naturally emerge from the histogram display as potentially important areas in the data space to explore. The event space is an emerging narrative composed of interwoven, partially decomposable threads in time. Sequences of events occur at multiple levels in a hierarchical theme space. These events are displayed against a backdrop of ongoing plans and expectations. The configuration of the event space, including the level and focus of the events, is dependent upon the perspective of the person who is viewing it. Finally, the event space visualization is founded upon a theoretical framework of event structures, including epochs that are defined by two landmark events, disrupting events, future and past events, continuous and discrete events, and three hierarchical levels in the event structure: Episodes, Events, and Elements.

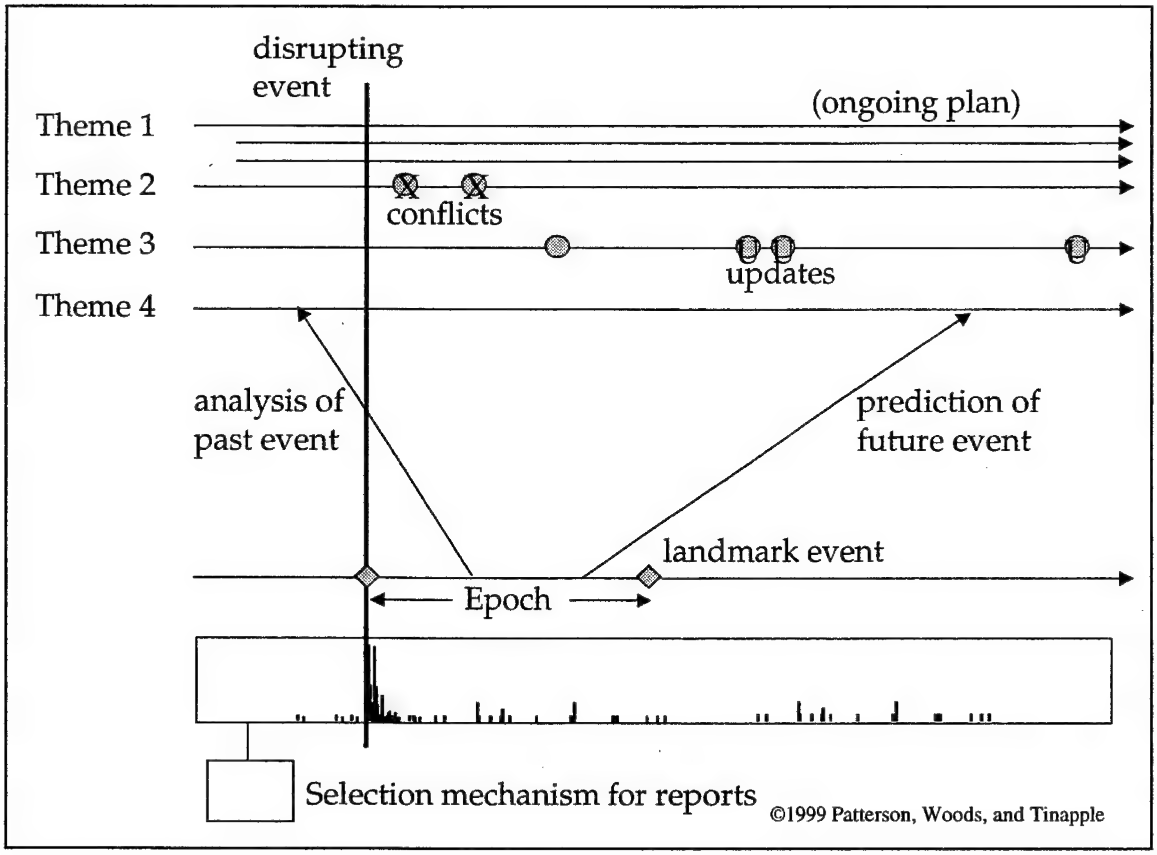


Figure 32. Overarching Workspace Concept: Visual Narratives

The Animock that instantiates this concept used a two-screen display (Figures 34 and 35). The Animock fly-through demonstration begins with the expert intelligence analyst who has been asked to perform a Quick Reaction Task (QRT) on the Ariane 501 accident, with which he is unfamiliar, in the process of selecting documents to browse. The analyst has already chosen three orthogonal ("AND") search facets: 1) Ariane (defined by the search terms Ariane, Arianespace, Ariane 5, OR Ariane V), 2) ESA (ESA, European Space Agency), and 3) lost (los*, fail*, explo*, destr*).[8] At this point, the analyst is adding to and removing facets to see how the changes impact the returned documents[9], displayed as A, B, C, etc. on the bottom of the interface as reports in time.[10] Then

[8] Note that this facet was created partly with computer suggestions about possible synonyms to add based on a computer dictionary and suggestions to truncate the verbs, which is represented by the "*" notation.

[9] This feature could make it easier to detect misspelled words in query formulations and investigate how adding keywords such as "1996" affect the results. One of the concerns about using 1996 as a keyword in the simulated task was whether or not that removed too many documents later in time that would have more definitive analyses of the accident.

[10] Note that a variation on this display would be for the entire report space to be displayed continuously on the bottom of the screen. We did not include this variation as in some cases there would be so many

the analyst creates the "Cluster" facet. A display pops up which naturally urges the analyst to type in alternative words to describe the concept such as payload, Cluster, and satellite.

At this point in the fly-through demonstration, the user tries two variations on a different kind of search facet aimed at improving the utility of the document set, rather than focusing on topicality as with the other facets. The "Aviation Week" node simply restricts documents to articles from the source Aviation Week and Space Technology. The "High Profit" node restricts the document set to ones with characteristics associated with high profit documents, including how long the document is, whether or not it is a summary or a full article, whether or not it is translated, and information regarding the document source[11]. Note that with this interface, the high profit or Aviation Week node could be connected to any of the topical nodes so that if a user wanted to apply the attribute selection to a larger set such as the entire database, it is easy to do so and quickly return to the other search configurations.

reports that the display would become a black area, but in cases where the data available is not large in relation to the returned sets, this could be a useful visualization.

[11] Note that this high profit node could be defined many ways. It could be a default setting in the software, it could be generated based on computer-inferred attributes of documents labeled "good" by the user, it could be created as customized settings by a group of analysts in a certain task domain, and it could be defined by the user with each analysis.

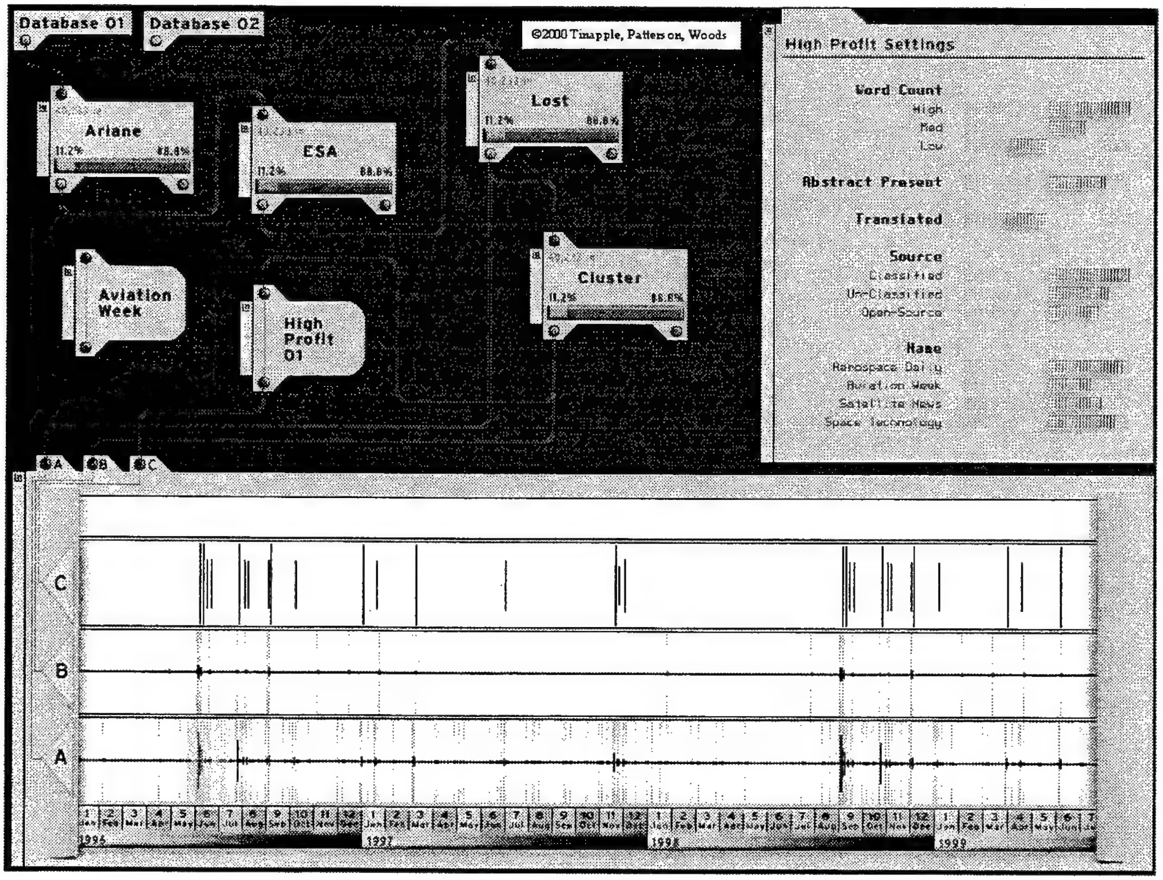


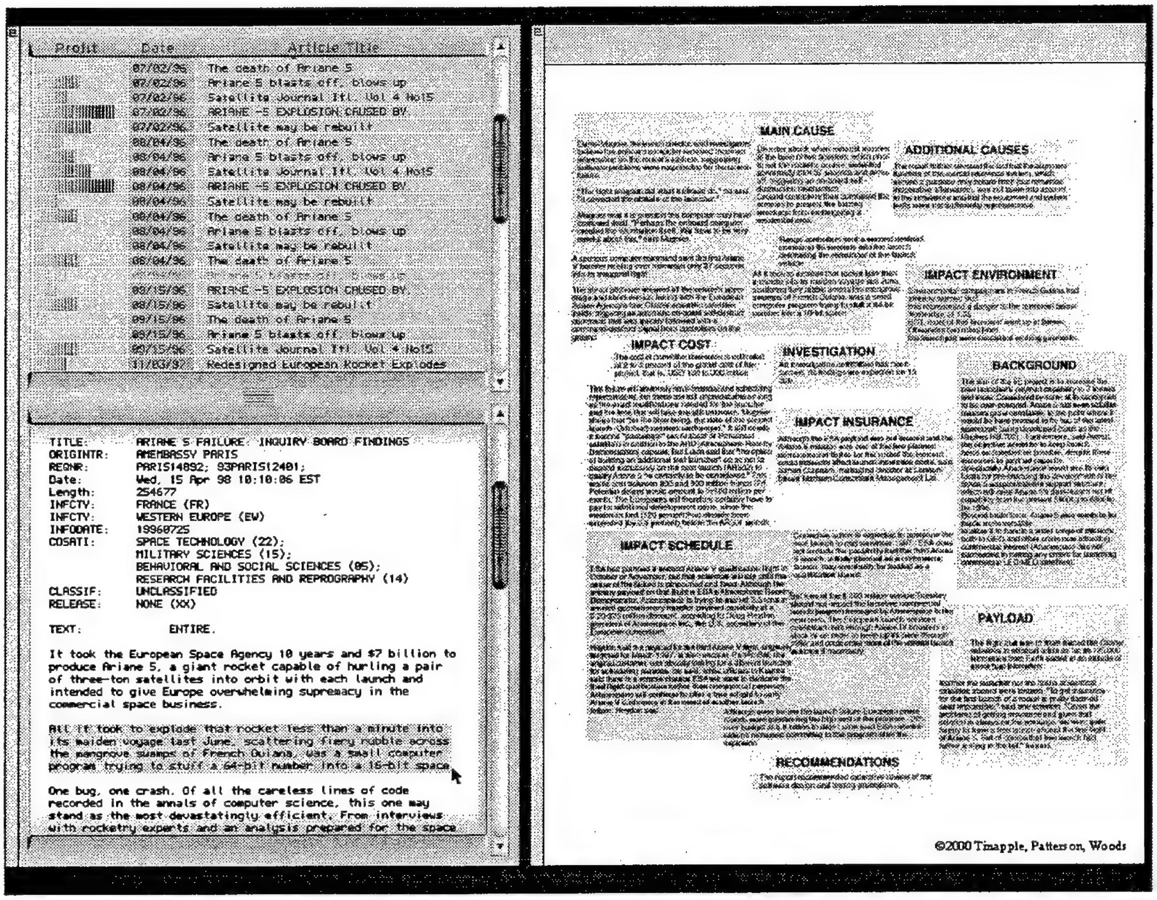
Figure 33. Left Screen of Integrated Workspace

The analyst then selects a portion of the selected documents in "A" by selecting a box on the query results visualization. These documents then appear in the browsing window on the right hand screen in the traditional browsing format by dates and titles.[12] The user then highlights the paragraph shown in Figure 35 and drags it to the right hand side of the screen, which we refer to as the "Snippet Workbench."[13] The workbench is designed so that the analyst can easily manipulate smaller units of information than documents in order to identify and resolve discrepant information. The labels on the text groupings are optionally entered by the user and serve as longshot landmarks when the mouse is used to naturally zoom in and out of the space. Other longshot landmarks could include

[12] Note that multiple browsers of this kind could be generated this way, and there are several options in designing the interface to support the navigation between these browsers using windows, tabs, or more complex integrated visualizations. We recommend, however, that the right hand side of the right screen not be tied to a particular browser window so as not to discourage sampling various areas of the report space.

[13] Note that this text selection strategy was observed with many of the study participants where the text was taken into a word processing program. Many of these participants referenced the selected text, such as by using footnotes, which was a labor-intensive process. In general, it was described that the work performed in the querying and browsing as well as the text manipulation was not easy to save as a unit for future reference.

information that is tagged with a bookmark or similar functionality to indicate a data conflict or unresolved issue that needs to be returned to prior to committing to an assessment.



Profile	Date	Article Title
	07/02/96	The death of Ariane 5
	07/02/96	Ariane 5 blasts off, blows up
	07/02/96	Satellite Journal Itl. Vol 4 No15
	07/02/96	ARIANE -5 EXPLOSION CAUSED BY
	07/02/96	Satellite may be rebuilt
	08/04/96	The death of Ariane 5
	08/04/96	Ariane 5 blasts off, blows up
	08/04/96	Satellite Journal Itl. Vol 4 No15
	08/04/96	ARIANE -5 EXPLOSION CAUSED BY
	08/04/96	Satellite may be rebuilt
	08/04/96	The death of Ariane 5
	08/04/96	Ariane 5 blasts off, blows up
	08/04/96	Satellite may be rebuilt
	08/04/96	The death of Ariane 5
	[illegible]	Ariane 5 blasts off, blows up
	08/15/96	ARIANE -5 EXPLOSION CAUSED BY
	08/15/96	Satellite may be rebuilt
	09/15/96	The death of Ariane 5
	08/15/96	Ariane 5 blasts off, blows up
	08/15/96	Satellite Journal Itl. Vol 4 No15
	11/03/97	Redesigned European Rocket Explodes

```
TITLE:        ARIANE 5 FAILURE: INQUIRY BOARD FINDINGS
ORIGINTR:     AMEMBASSY PARIS
REQNR:        PARIS14892; 93PARIS12401;
Date:         Wed, 15 Apr 98 10:10:06 EST
Length:       254677
INFCTY:       FRANCE (FR)
INFCTY:       WESTERN EUROPE (EW)
INFODATE:     19960725
COSATI:       SPACE TECHNOLOGY (22);
              MILITARY SCIENCES (15);
              BEHAVIORAL AND SOCIAL SCIENCES (05);
              RESEARCH FACILITIES AND REPROGRAPHY (14)
CLASSIF:      UNCLASSIFIED
RELEASE:      NONE (XX)

TEXT:              ENTIRE.
```

It took the European Space Agency 10 years and $7 billion to produce Ariane 5, a giant rocket capable of hurling a pair of three-ton satellites into orbit with each launch and intended to give Europe overwhelming supremacy in the commercial space business.

All it took to explode that rocket less than a minute into its maiden voyage last June, scattering fiery rubble across the mangrove swamps of French Guiana, was a small computer program trying to stuff a 64-bit number into a 16-bit space.

One bug, one crash. Of all the careless lines of code recorded in the annals of computer science, this one may stand as the most devastatingly efficient. From interviews with rocketry experts and an analysis prepared for the space

Figure 34. Right Screen of Integrated Workspace

In the final portion of the Animock fly-through, we replace the right hand screen with a "Themespace" visualization (Figure 31). This visualization is intended to provide a longshot that encourages meta-level reflection on the analysis process. The goal is to be able to see holes in the document selection, snippet selection, conflict resolution, and thematic story generation processes.[14] In addition, the longshot is built upon the smaller data elements in the other views, and so can be used to re-organize them at higher levels of abstraction. For example, the labels on the text groupings could be reorganized in order to move closer towards the logical sequence in a briefing and new groupings could be created that would

[14] Note that this could be accomplished through visualizations that allow analysts to do so as well as by adding computer intelligence that could actively critique the process when the visualization is called up. A particularly useful critique at this time would include suggestions about potential updates that could overturn the analysis on sub-themes.

need to be filled in at a later time. At this point, we ended the Animock in order to encourage discussion about what would be needed in the Themescape visualization to be useful because it is not a tool that currently exists for them. In addition, we wanted to know how and when the Themescape idea could help analysts in finding further relevant and high quality information from the left screen, in order to reduce the risks of prematurely closing the analysis process.

PART V. SUMMARY

In summary, we detailed a complete, beginning-to-end Cognitive Systems Engineering (CSE) project tackling the challenges of conducting intelligence analysis under the condition of data overload that:

- revealed the world of the intelligence analyst and grounded general concepts to the particulars of the situation the professional analyst faces,
- identified characteristics of intelligence analysis that were similar and unique to other settings,
- contributed to our general understanding of data overload,
- generated practice-centered criteria for evaluating proposed solutions to data overload,
- served as a basis for interaction and as a stimulus to a more constructive dialogue across analysts, developers and others for useful design directions to pursue, and
- generated nine modular "design seeds" that represent innovative directions for solutions to data overload that could be incorporated into short- and long-term design efforts as well as be used in follow-on research to improve the usefulness of the design seeds, and
- instantiated the design seeds in an overarching Visual Narratives concept and illustrated the concept using an animated fly-through, or Animock, based on an analyst tasked with the Ariane 501 scenario.

This project also provided methodological contributions. This case study serves as example of a Cognitive Systems Engineering approach that:

- views science and design as complementary, mutually reinforcing activities,
- illustrates design as an iterative "bootstrap" process,
- uses prototypes as tools for discovery to probe the interaction of people, technology and work,
- separates out learning on three levels throughout the design process: understanding the challenges in a domain, determining what would be useful aids to domain practitioners, and improving the usability of artifacts, and
- focuses our Research & Development investments on the "usefulness" of designs in order to target leverage points that will have the most impact on the end practitioners' ability to meet domain challenges.

We believe that a critical success criterion in a Cognitive Systems Engineering project is for the domain practitioners, as the main problem holders, to recognize the problems attempting to be addressed as their own. We provide the following testimonial as evidence of this problem recognition by a stakeholder in the intelligence analysis community:

> "This work is very crucial to the Air Force and the intelligence community, in particular, which is faced with increasing data flows and a declining work force...We need some way to preserve the long-term analyst's wisdom and/or experiences...This is exciting work with real promise."
>
> -- Intelligence analyst

We believe that it is only by meeting this criteria that R&D projects will result in system designs that will ultimately prove useful in providing effective computerized support to help them address the difficult challenges of conducting inferential analysis under data overload.

ACKNOWLEDGMENTS

Funding was provided by the Human Effectiveness Directorate of the Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio, with truly exceptional support and guidance from Gil Kuperman. Additional funding was provided by a National Science Foundation Graduate Fellowship. Any opinions, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. Many of these ideas have roots in stimulating discussions over many years with collaborators, some of which include Robert Hoffman, Jane Malin, Klaus Christoffersen, James Corban, Richard Cook, Gary Klein, Wayne Redenbarger, Dan Howland, and Michael McNeese. Janet Reynolds greatly facilitated the project, particularly with the study. Many intelligence analysts generously dedicated their time and expertise throughout the project. In particular Jay Finley, James Homer, and Mickey Wilkinson provided excellent guidance and feedback.

REFERENCES

Bates, M.J. (1979). Information Search Tactics. *Journal of the American Society for Information Science,* 30: 205-214.

Billings, C. E. (1996). *Aviation Automation: The search for a human-centered approach.* Hillsdale, NJ: Erlbaum.

Blair, D.C. (1980). Searching biases in large interactive document retrieval systems. *Journal of the American Society for Information Science,* 31, 271-277.

Blair, D. C. & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system, *Communications of the ACM* 28, 3, 289-299.

Bower G. & Morrow, D. G. (1990). Mental models in narrative comprehension. *Science, 24,* 44-48.

Cook, R. I., & Woods, D. D. (1996). Adapting to new technology in the operating room. *Human Factors, 38(4),* 593-613.

Flach, J., Hancock, P., Caird, J. & Vicente, K. (Eds.) (1995). *An Ecological Approach To Human Machine Systems I: A Global Perspective.* Hillsdale, NJ: Lawrence Erlbaum.

Guerlain, S., Smith, P.J., Obradovich, J. H., Rudmann, S., Strohm, P., Smith, J.W., Svirbely, J., & Sachs, L. (1999). Interactive Critiquing as a Form of Decision Support: An Empirical Evaluation. *Human Factors, 41*(1), 72-89.

Hutchins, E. (1995). *Cognition in the Wild.* Cambridge, MA: MIT Press.

James, W. (1981). *The Principles of Psychology.* New York, H. Holt and Company [original 1918].

Johannesen, L.J., Cook, R.I., & Woods, D.D. (1994). Cooperative communications in dynamic fault management. *Proceedings of the 38th Annual Meeting of the Human Factors and Ergonomics Society.* Nashville, TN.

Josephson, J., & Josephson, S. (1994). *Abductive Inference.* New York, NY: Cambridge University Press.

Malin, J., Schreckenghost, D., Woods, D., Potter, S., Johannesen, L., Holloway, M., & Forbus, K. (1991). *Making Intelligent Systems Team Players: Case Studies and Design Issues.* NASA Technical Memo 104738.

Morse, E., & Lewis, M. (1997). Why information retrieval visualizations sometimes fail. In *Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 12-15, 1997, pp. 1680 - 1685.

Olsen K.A., Sochats K.M., & Williams, J.G. (1998). Full text searching and information overload. *International Information and Library Review* 30: (2) 105-122.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology, 46,* 461-488.

Patterson, E.S., Roth, E.M., & Woods, D.D. (1999). Aiding the intelligence analyst in situations of data overload: a simulation study of computer-supported inferential analysis under data overload. *Institute for Ergonomics/Cognitive Systems Engineering Laboratory Report, ERGO-CSEL 99-TR-02,* The Ohio State University, Columbus OH.

Rabbitt, P. (1984). The control of attention in visual search. In R. Parasuraman and D. R. Davies (Eds.), *Varieties of Attention.* New York: Academic Press.

Rasmussen, J. (1985). Trends in human reliability analysis. *Ergonomics, 28*(8), 1185-1196.

Rasmussen, J., Pejtersen, A. M., & Goldstein, L. P. (1994). *Cognitive Systems Engineering.* New York: John Wiley and Sons

Roth, E. M., Woods D. D., & Pople, H. E. Jr. (1992). Cognitive simulation as a tool for cognitive task analysis. *Ergonomics, 35,* 1163–1198.

Saracevic, T., Kantor, P., Chamis, A.Y., & Trivison, D. (1988). A study of information seeking and retrieving (3 parts). *Journal of the American Society for Information Science,* 39, 161-216.

Sarter, N.B., & Woods, D.D. (1992) "Pilot Interaction with Cockpit Automation: Operational Experiences with the Flight Management System (FMS)." *International Journal of Aviation Psychology, 2(4),* 303-321.

Sarter, N., Woods, D. D. & Billings, C. (1997). Automation surprises. In G. Salvendy, (Ed.) *Handbook of Human Factors/Ergonomics* (2nd ed.). New York: John Wiley and Sons.

Schum, D.A. (1994). *The Evidential Foundations of Probabilistic Reasoning.* New York: John Wiley and Sons.

Shattuck, L. & D.D. Woods. (1997) Communication Of Intent In Distributed Supervisory Control Systems. In Proceedings of the 41st Annual Meeting of the Human Factors and Ergonomics Society, September.

Shute, S. J., & Smith, P. J. (1992). Knowledge-based search tactics. *Information Processing & Management* 29(1), 29-45.

Thronesbery, C., Christoffersen, K., & Malin, J. (1999). Situation-oriented displays of space shuttle data. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*, September 27 – October 1, Houston, Texas, 284-288.

Tukey, J.W. (1977). *Exploring Data Analysis*. Reading Massachusetts: Addison-Wesley.

Vicente, K. J. & Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. *IEEE Transactions on Systems, Man, and Cybernetics, 22(4)*, 589-606.

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1996) Visualizing the non-visual: Spatial analysis and interaction with information from text documents. *Proceedings of Info Viz 96.*

Wolfe, J. M. (1992). The parallel guidance of visual attention. *Current Directions in Psychological Science, 1*, 124-128.

Woods, D. D. (1984). Visual momentum: A concept to improve the cognitive coupling of person and computer. *International Journal of Man-Machine Studies, 21*, 229-244.

Woods, D. D. (1994). Cognitive demands and activities in dynamic fault management: abductive reasoning and disturbance management. In N. Stanton (Eds.), *Human factors in alarm design* Bristol, PA: Taylor and Francis.

Woods, D. D. (1995a). The alarm problem and directed attention in dynamic fault management. *Ergonomics, 38*(11), 2371-2393.

Woods, D. D. (1995b). Towards a theoretical base for representation design in the computer medium: ecological perception and aiding human cognition. In J. Flach, P. Hancock, J. Caird, and K. Vicente, (Eds.) *An Ecological Approach To Human Machine Systems I: A Global Perspective*, Hillsdale, NJ: Lawrence Erlbaum.

Woods, D.D. (1998). Designs are Hypotheses about How Artifacts Shape Cognition and Collaboration. *Ergonomics*, 41, 168-173.

Woods, D.D., Patterson, E.S., & Roth, E.M. (1998). Aiding the intelligence analyst in situations of data overload: a diagnosis of data overload. *Institute for Ergonomics/Cognitive Systems Engineering Laboratory Report, ERGO-CSEL 98-TR-03,* The Ohio State University, Columbus OH.

Woods, D. D., Pople, H.E. Jr., & Roth, E. M. (1990). *The Cognitive Environment Simulation as a Tool for Modeling Human Performance and Reliability, Volumes I and II* (Technical Report NUREG-CR-5213). Washington D.C.: U.S. Nuclear Regulatory Commission.

Woods, D. D. & Sarter, N. B. (1993). Evaluating the impact of new technology on human-machine cooperation. In Wise, Hopkin & Stager, (Eds.) *Verification and Validation of Human-Machine Systems*, Springer-Verlag, 1993.

Woods, D.D. & Sarter, N.B. (2000). Learning from Automation Surprises and "Going Sour" Accidents. In N.B. Sarter and R. Amalberti (Eds.), *Cognitive Engineering in the Aviation Domain* (pp. 327-353). Hillsdale, NJ: LEA.

Woods, D. D. & Watts, J. C. (1997) How not to have to navigate through too many displays. In M. Helander (Ed.), *Handbook of Human-Computer Interaction, 2nd edition*. North-Holland: Elsevier Science Publishers, B.V.